

# Convergence of latent mixing measures in nonparametric and mixture models<sup>1</sup>

XuanLong Nguyen  
xuanlong@umich.edu  
Department of Statistics  
University of Michigan

## Abstract

We consider Wasserstein distance functionals for assessing the convergence of latent discrete measures, which serve as mixing distributions in hierarchical and nonparametric mixture models. We clarify the relationships between Wasserstein distances of mixing distributions and  $f$ -divergence functionals such as Hellinger and Kullback-Leibler distances on the space of mixture distributions using various identifiability conditions. The convergence in Wasserstein metrics for discrete measures has a natural interpretation of the convergence of individual atoms that provide support for the discrete measure. It is also typically stronger than the weak convergence induced by standard  $f$ -divergence metrics. We establish rates of convergence of posterior distributions for latent discrete measures in several mixture models, including finite mixtures of multivariate distributions, finite mixtures of Gaussian processes and infinite mixtures based on the Dirichlet process.

## 1 Introduction

A notable feature in the development of hierarchical and Bayesian nonparametric models is the role of discrete probability measures, which serve as mixing distributions to combine relatively simple models into richer classes of statistical models (Lindsay, 1995; McLachlan and Basford, 1988). In recent years the mixture modeling methodology has been significantly extended, by many authors taking the mixing measure to be random and infinite dimensional via suitable priors constructed in a nested, hierarchical and nonparametric manner. This results in rich models that can fit more complex and high dimensional data (see, e.g., Gelfand et al. (2005); Teh et al. (2006); Rodriguez et al. (2008); Petrone et al. (2009); Nguyen (2010) for several examples of such models, as well as a recent book Hjort et al. (2010)).

The focus of this paper is to analyze convergence behavior of the posterior distribution of latent mixing measures as they arise in several mixture models, including infinite mixture and other nonparametric models. Let  $G = \sum_{i=1}^k p_i \delta_{\theta_i}$  denote a discrete probability measure. Atoms  $\theta_i$ 's are elements in space  $\Theta$ , while vector of probabilities  $\mathbf{p} = (p_1, \dots, p_k)$

---

<sup>1</sup> AMS 2000 subject classification. Primary 62F15, 62G05; secondary 62G20.

Key words and phrases: Mixture distributions, hierarchical models, Bayesian nonparametrics, Wasserstein distance,  $f$ -divergence, rates of convergence, Dirichlet processes, Gaussian processes.

lies in a  $k - 1$  dimensional probability simplex. In a mixture setting,  $G$  is combined with a likelihood density  $f(\cdot|\theta)$  with respect to a dominating measure  $\mu$  on  $\mathcal{X}$ , to yield the mixture density:  $p_G(x) = \int f(x|\theta)dG(\theta) = \sum_{i=1}^k p_i f(x|\theta_i)$ . In a clustering application, atoms  $\theta_i$ 's represent distinct behaviors in a heterogeneous data population, while mixing probabilities  $p_i$ 's are the associated proportions of such behaviors. Under this interpretation, there is a need for comparing and assessing the quality of the discrete measure  $\hat{G}$  estimated on the basis of available data. An important work in this direction is by Chen (1995), who used the  $L_1$  metric on the cumulative distribution functions on the real line to study convergence rates of the mixing measure  $G$ . Building upon Chen's work, Ishwaran, James and Sun (2001) established the posterior consistency of a finite dimensional Dirichlet prior for Bayesian mixture models. Their analysis is specific to univariate finite mixture models, with  $k$  bounded by a known constant, while our interest is when  $k$  may be unbounded and/or  $\Theta$  has high or infinite dimensions. For instance,  $\Theta$  may be a subset of a function space as in the work of Gelfand et al. (2005); Nguyen (2010), or a space of probability measures (Rodriguez et al., 2008).

The analysis of consistency and convergence rates of posterior distributions for Bayesian estimation have seen much progress in the past decade. Key recent references include Barron et al. (1999); Ghosal et al. (2000); Shen and Wasserman (2001); Walker (2004); Ghosal and van der Vaart (2007a); Walker et al. (2007). Analysis of specific mixture models in a Bayesian setting have also been studied extensively Ghosal et al. (1999); Genovese and Wasserman (2000); Ishwaran and Zarepour (2002); Ghosal and van der Vaart (2007b). All these work primarily focus on the convergence in the topology of Hellinger or a comparable distance metric in the space of data densities  $p_G$ . On the other hand, there are far less work concerning with the convergence behavior of latent mixing measures  $G$ . Notably, the analysis of convergence for mixing (smooth) densities often arised in the context of frequentist estimation for deconvolution problems, mainly with kernel density estimation methods Zhang (1990); Fan (1991).

The primary contribution of this paper is to show that the Wasserstein distances provide a natural and useful metric for the analysis of convergence for latent and discrete mixing measures in mixture models, and to establish convergence rates of posterior distributions in a number of well-known Bayesian nonparametric and mixture models. Wasserstein distances originally arised in the problem of optimal transportation Villani (2003). It has been utilized in a number of statistical contexts (e.g., Dudley (1976); Mallows (1972); Bickel and Freedman (1981); del Barrio et al. (1999)). For discrete probability measures, they can be obtained by a minimum matching (or moving) procedure between the sets of atoms that provide support for the measures under comparison, and consequently are simple to compute. Suppose that  $\Theta$  is equipped with a metric  $\rho$ . Let  $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$ . Then, the  $L_r$  Wasserstein distance metric on the space of discrete probability measures with support in  $\Theta$ , namely,  $\bar{\mathcal{G}}(\Theta)$ , is:

$$d_\rho(G, G') = \left[ \inf_{\mathbf{q}} \sum_{i,j} q_{ij} \rho^r(\theta_i, \theta'_j) \right]^{1/r},$$

where the infimum is taken over all joint probability distributions on  $[1, \dots, k] \times [1, \dots, k']$  such that  $\sum_j q_{ij} = p_i$  and  $\sum_i q_{ij} = p'_j$ .

As clearly seen from this definition, the Wasserstein distances inherit directly the metric of the space of atomic support  $\Theta$ , suggesting that they can be useful for assessing estimation procedures for discrete measures in hierarchical models. It is worth noting that if  $(G_n)_{n \geq 1}$  is a sequence of discrete probability measures with  $k$  distinct atoms and  $G_n$  tends to some  $G_0$  in  $d_\rho$  metric, then  $G_n$ 's ordered set of atoms must converge to  $G_0$ 's atoms in  $\rho$  after some permutation of atom labels. Thus, in the clustering application illustrated above, convergence of mixing measure  $G$  may be interpreted as the convergence of distinct typical behavior  $\theta_i$ 's that characterize the heterogeneous data population. A hint for the relevance of the Wasserstein distances can be drawn from an observation that the  $L_1$  distance for the CDF's of univariate random variables, as studied by Chen (1995), is in fact a special case of the  $L_1$  Wasserstein metric when  $\Theta = \mathbb{R}$ .

The plan for the paper is as follows. Section 2 explores the relationship between Wasserstein distances for mixing measures and well-known divergence functionals for mixture densities in a mixture model. We produce a simple lemma which gives an upper bound of  $f$ -divergences between mixture densities by certain Wasserstein distances between mixing measures. This implies that  $d_\rho$  topology can be stronger than those induced by divergences between mixture densities. Next, we consider various identifiability conditions under which convergence of mixture densities entails convergence of mixing measures in the Wasserstein metric. We present two theorems, which provide upper bounds of  $d_\rho(G, G')$  in terms of divergences between  $p_G$  and  $p_{G'}$ . Theorem 1 is applicable to mixing measures with a bounded number of atomic support, generalizing a result from Chen (1995). Theorem 2 is applicable to mixing measures with unbounded number of support points, but is restricted to only convolution mixture models.

Section 3 focuses on the convergence of posterior distributions of latent mixing measures in a Bayesian nonparametric setting. Here, the mixing measure  $G$  is endowed with a prior distribution  $\Pi$ . Assuming an  $n$ -sample  $X_1, \dots, X_n$  that is generated according to  $p_{G_0}$ , we study conditions under which the posterior distribution of  $G$ , namely,  $\Pi(\cdot | X_1, \dots, X_n)$ , contracts to the "truth"  $G_0$  under the  $d_\rho$  metric, and provide the contraction rates. In Theorems 3 and 4 of Section 3, we establish the convergence rates for the posterior distribution for  $G$  in terms of  $d_\rho$  metric. These results are proved using the standard approach of Ghosal, Ghosh and van der Vaart (2000). Our convergence theorems have several notable features. They rely on separate conditions for the prior  $\Pi$  and likelihood function  $f$ , which are typically simpler to verify than conditions formulated in terms of mixture densities. The claim of convergence in Wasserstein metrics is typically stronger than the weak convergence induced by the Hellinger metric in the existing work mentioned above.

In Section 4 posterior consistency and convergence rates of latent mixing measures are derived for a number of well-known mixture models in the literature, including finite mixtures of multivariate distributions, infinite mixtures based on Dirichlet processes, and finite mixtures of Gaussian processes. For finite mixtures with bounded number of atomic support, the posterior convergence rate for mixing measure is the minimax optimal  $n^{-1/4}$

under suitable identifiability conditions. For Dirichlet process mixtures defined on  $\mathbb{R}^d$ , specific rates are established under smoothness conditions of the likelihood density function  $f$ . In particular, for ordinary smooth likelihood densities with smoothness  $\beta$  (e.g., Laplace), the rate achieved is  $(\log n/n)^\gamma$  for any  $\gamma < \frac{2}{(d+2)(4+(2\beta+1)d)}$ . For supersmooth likelihood densities with smoothness  $\beta$  (e.g., normal), the rate achieved is  $(\log n)^{-1/\beta}$ . Finally, for finite mixtures of Gaussian processes, we are also able to establish a convergence rate by utilizing a result on (single) Gaussian process prior by van der Vaart and van Zanten van der Vaart and van Zanten (2008b).

**Notations.** For ease of notations, we also use  $f_i$  in place of  $f(\cdot|\theta_i)$ , and  $f'_j$  in place of  $f(\cdot|\theta'_j)$  for likelihood density functions. Divergences studied in the paper include the total variational distance:  $d_V(p_G, p_{G'}) := \frac{1}{2} \int |p_G(x) - p_{G'}(x)| d\mu$ , the Hellinger distance:  $d_h^2(p_G, p_{G'}) := \frac{1}{2} \int (\sqrt{p_G(x)} - \sqrt{p_{G'}(x)})^2 d\mu$ , and the Kullback-Leibler divergence:  $d_K(p_G, p_{G'}) = \int p_G(x) \log(p_G(x)/p_{G'}(x)) d\mu$ . These divergences are related by  $d_V^2/2 \leq d_h^2 \leq d_V$  and  $d_h^2 \leq d_K/2$ .  $N(\epsilon, \Theta, \rho)$  denotes the covering number of the metric space  $(\Theta, \rho)$ , i.e., the minimum number of  $\epsilon$ -balls needed to cover the entire space  $\Theta$ .  $D(\epsilon, \Theta, \rho)$  denotes the packing number of  $(\Theta, \rho)$ , i.e., the maximum number of points that are mutually separated by at least  $\epsilon$  in distance. They are related by  $N(\epsilon, \Theta, \rho) \leq D(\epsilon, \Theta, \rho) \leq N(\epsilon/2, \Theta, \rho)$ .  $\text{Diam}(\Theta)$  denotes the diameter of  $\Theta$ .

## 2 Wasserstein distances for mixing measures

### 2.1 Definition and a basic inequality

Let  $(\Theta, \rho)$  be a space equipped with a non-negative distance function  $\rho$  such that  $\rho(\theta_1, \theta_2) = 0$  if and only if  $\theta_1 = \theta_2$ . If in addition,  $\rho$  is symmetric ( $\rho(\theta_1, \theta_2) = \rho(\theta_2, \theta_1)$ ), and satisfies the triangle inequality, then it is a proper metric. A discrete probability measure  $G$  on a measure space equipped with the Borel sigma algebra takes the form  $G = \sum_{i=1}^k p_i \delta_{\theta_i}$  for some  $k \in \mathbb{N} \cup \{+\infty\}$ , where  $\mathbf{p} = (p_1, p_2, \dots, p_k)$  denotes the proportion vector, while  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$  are the associated atoms in  $\Theta$ .  $\mathbf{p}$  has to satisfy  $0 \leq p_i \leq 1$  and  $\sum_{i=1}^k p_k = 1$ . Likewise,  $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$  is another discrete probability measure that has at most  $k'$  distinct atoms. Let  $\mathcal{G}_k(\Theta)$  denote the space of all discrete probability measures with at most  $k$  atoms. Let  $\mathcal{G}(\Theta) = \cup_{k \geq 1} \mathcal{G}_k(\Theta)$ , the set of all discrete measures with finite support. Finally,  $\bar{\mathcal{G}}(\Theta)$  denotes the space of all discrete measures (including those with countably infinite support).

Let  $\mathbf{q} = (q_{ij})_{i \leq k; j \leq k'} \in [0, 1]^{k \times k'}$  denote a joint probability distribution on  $\mathbb{N}_+ \times \mathbb{N}_+$  that satisfies the marginal constraints:  $\sum_{i=1}^k q_{ij} = p'_j$  and  $\sum_{j=1}^{k'} q_{ij} = p_i$  for any  $i = 1, \dots, k; j = 1, \dots, k'$ . Let  $\mathcal{Q}(\mathbf{p}, \mathbf{p}')$  denote the space of all such joint distributions. We start with the  $L_1$  Wasserstein distance:

**Definition 1.** Let  $\rho$  be a distance function on  $\Theta$ . The Wasserstein distance functional for two discrete measures  $G(\mathbf{p}, \boldsymbol{\theta})$  and  $G'(\mathbf{p}', \boldsymbol{\theta}')$  is:

$$d_\rho(G, G') = \inf_{\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}')} \sum_{i,j} q_{ij} \rho(\theta_i, \theta'_j). \quad (1)$$

We focus mainly on the  $L_1$  Wasserstein  $d_\rho$ , and  $L_2$  Wasserstein distance. The latter corresponds to the square root of  $d_{\rho^2}$  in our definition, where  $\rho(\theta_i, \theta'_j)$  is replaced by  $\rho^2(\theta_i, \theta'_j)$ . Note that  $d_\rho^2(G, G') \leq d_{\rho^2}(G, G')$  by an application of Cauchy-Schwarz inequality. We will consider a variety of choices of distance  $\rho$  in the sequel.

From here on, discrete measure  $G \in \bar{\mathcal{G}}(\Theta)$  plays the role of the mixing distribution in a mixture model. Let  $f(x|\theta)$  denote the density (with respect to a dominating measure  $\mu$ ) of a random variable  $X$  taking value in  $\mathcal{X}$ , given parameter  $\theta \in \Theta$ . For the ease of notations, we also use  $f_i(x)$  for  $f(x|\theta_i)$ . Combining  $G$  with the likelihood function  $f$  yields a mixture distribution for  $X$  that takes the following density:

$$p_G(x) = \int f(x|\theta) dG(\theta) = \sum_{i=1}^k p_i f_i(x).$$

First we state a general result that relates Wasserstein distances between mixing measures  $G, G'$ , namely,  $d_\rho(G, G')$  and divergences between mixture densities  $p_G, p_{G'}$ . Divergence functionals that play important role in this paper are the total variational distance, Hellinger distance, and the Kullback-Leibler distance. All these are in fact instances of a broader class of divergence functionals known as the  $f$ -divergences (Csizar, 1966; Ali & Silvey, 1967):

**Definition 2.** Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  denote a convex function. An  $f$ -divergence (or Ali-Silvey distance) between two probability densities  $f_i$  and  $f'_j$  is defined as  $d_\phi(f_i, f'_j) = \int \phi(f'_j/f_i) f_i d\mu$ . Likewise, the  $f$ -divergence between  $p_G$  and  $p_{G'}$  is  $d_\phi(p_G, p_{G'}) = \int \phi(p_{G'}/p_G) p_G d\mu$ .

For  $\phi(u) = \frac{1}{2}(\sqrt{u}-1)^2$  we obtain the squared Hellinger. For  $\phi(u) = \frac{1}{2}|u-1|$  we obtain the variational distance. For  $\phi(u) = -\log u$ , we obtain the Kullback-Leiber divergence.  $f$ -divergence functionals can be used as a distance function or metric on  $\Theta$ , motivating the following definition.

**Definition 3.** When  $\rho$  is taken to be an  $f$ -divergence,  $\rho(\theta_i, \theta'_j) = d_\phi(f_i, f'_j)$ , for a convex function  $\phi$ , the corresponding Wasserstein distance functional is called a composite Wasserstein distance:

$$d_{\rho\phi}(G, G') = \inf_{\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}')} \sum_{i,j} q_{ij} d_\phi(f_i, f'_j).$$

In particular,  $d_V, d_h, d_K$  induce the composite Wasserstein distances  $d_{\rho V}, d_{\rho h}, d_{\rho K}$ , respectively. Let  $d_{\rho h^2}(G, G')$  denote the composite Wasserstein distance function obtained by taking  $\rho(\theta_i, \theta'_j) := d_h^2(f_i, f'_j)$ . The following result shows that any  $f$ -divergence between mixture distribution  $p_G, p_{G'}$  is dominated by a Wasserstein distance for a suitable choice of distance  $\rho$ .

**Lemma 1.** *Let  $G, G' \in \bar{\mathcal{G}}(\Theta)$  such that both  $d_\phi(p_G, p_{G'})$  and  $d_{\rho\phi}(G, G')$  are finite for some convex function  $\phi$ . Then,  $d_\phi(p_G, p_{G'}) \leq d_{\rho\phi}(G, G')$ .*

As will be evident in the sequel, this lemma is also handy in enabling us to obtain lower bounds on small ball probabilities in the space of mixture densities  $p_G$ , in terms of small ball probabilities in the metric space  $(\Theta, \rho)$ . The latter quantities are typically easier to obtain estimates than the former.

**Example 1.** Suppose that  $\Theta = \mathbb{R}^d$ ,  $\rho$  is the Euclidean metric,  $f(x|\theta)$  is the multivariate normal density  $N(\theta, I_{d \times d})$  with mean  $\theta$  and identity covariance matrix, then  $d_h^2(f_i, f'_j) = 1 - \exp -\frac{1}{8}\|\theta_i - \theta'_j\|^2 \leq \frac{1}{8}\|\theta_i - \theta'_j\|^2$ . So,  $d_{\rho h^2}(G, G') \leq d_{\rho^2}(G, G')/8$ . The above lemma then entails that  $d_{h^2}(p_G, p_{G'}) \leq d_{\rho^2}(G, G')/8$ .

Similarly, for the Kullback-Leibler divergence, since  $d_K(f_i, f'_j) = \frac{1}{2}\|\theta_i - \theta'_j\|^2$ , by Lemma 1,  $d_K(p_G, p_{G'}) \leq d_{\rho K}(G, G') = \frac{1}{2}d_{\rho^2}(G, G')$ . Next, suppose that  $\Theta$  is a compact subset of  $\mathbb{R}^d$  and consider  $\phi(u) = (\log u)^2$ , which is a convex function on  $[0, \infty)$ . We have  $\int f_i(\log f_i/f'_j)^2 = O(\|\theta_i - \theta'_j\|^2)$ , so  $\int p_G[\log(p_G/p_{G'})]^2 \leq O(d_{\rho^2}(G, G'))$ .

For another example, if  $f(x|\theta)$  is a Gamma density with location parameter  $\theta$ ,  $\Theta$  is a compact subset in  $\mathbb{R}$ . Then  $d_K(f_i, f'_j) = O(|\theta_i - \theta_j|)$ . This entails that  $d_K(p_G, p_{G'}) \leq O(d_\rho(G, G'))$ .

## 2.2 Wasserstein metric identifiability in finite mixture models

Lemma 1 shows that for many choices of  $\rho$ ,  $d_\rho$  yields a stronger topology on  $\bar{\mathcal{G}}(\Theta)$  than the topology induced by  $f$ -divergences on the space of mixture distributions  $p_G$ . In other words, convergence of  $p_G$  may not imply convergence of  $G$  in  $d_\rho$  metric. To ensure this property, additional conditions are needed on the space of discrete measures  $\bar{\mathcal{G}}(\Theta)$ , along with identifiability conditions for the family of likelihood functions  $\{f(\cdot|\theta), \theta \in \Theta\}$ .

The classical definition of Teicher Teicher (1961) specifies the family  $\{f(\cdot|\theta), \theta \in \Theta\}$  to be identifiable if for any  $G, G' \in \mathcal{G}(\Theta)$ ,  $\|p_G - p_{G'}\|_\infty = 0$  implies that  $G = G'$ . We need a slightly stronger version, allowing for the inclusion for discrete measures with infinite support:

**Definition 4.** *The family  $\{f(\cdot|\theta), \theta \in \Theta\}$  is finitely identifiable if for any  $G \in \mathcal{G}_\Theta$  and  $G' \in \bar{\mathcal{G}}_\Theta$ ,  $|p_G(x) - p_{G'}(x)| = 0$  for almost all  $x \in \mathcal{X}$  implies that  $G = G'$ .*

To obtain convergence rates, we also need the notion of strong identifiability of Chen (1995), herein adapted to a multivariate setting.

**Definition 5.** *Assume that  $\Theta \subseteq \mathbb{R}^d$  and  $\rho$  is the Euclidean metric. The family  $\{f(\cdot|\theta), \theta \in \Theta\}$  is strongly identifiable if  $f(x|\theta)$  is twice differentiable in  $\theta$  and for any finite  $k$  and  $k$  different  $\theta_1, \dots, \theta_k$ , the equality*

$$\text{ess sup}_{x \in \mathcal{X}} \left| \sum_{i=1}^k \alpha_i f(x|\theta_i) + \beta_i^T Df(x|\theta_i) + \gamma_i^T D^2 f(x|\theta_i) \gamma_i \right| = 0 \quad (2)$$

implies that  $\alpha_i = 0$ ,  $\beta_i = \gamma_i = \mathbf{0} \in \mathbb{R}^d$  for  $i = 1, \dots, k$ . Here, for each  $x$ ,  $Df(x|\theta_i)$  and  $D^2f(x|\theta_i)$  denote the gradient and the Hessian at  $\theta_i$  of function  $f(x|\cdot)$ , respectively.

Finite identifiability is satisfied for the family of Gaussian distributions (Teicher, 1960), see also Theorem 1 of Ishwaran and Zarepour (2002). Chen identified a broad class of families, including the Gaussian family, for which the strong identifiability condition holds (Chen, 1995).

Define  $\psi(G, G') = \sup_x |p_G(x) - p_{G'}(x)|/d_{\rho^2}(G, G')$  if  $G \neq G'$  and  $\infty$  otherwise. Also define  $\psi_1(G, G') = d_V(p_G, p_{G'})/d_{\rho^2}(G, G')$  if  $G \neq G'$  and  $\infty$  otherwise. The notion of strong identifiability is useful via the following key result, which generalizes Chen's result to  $\Theta$  of arbitrary dimensions.

**Theorem 1. (Strong identifiability).** *Suppose that  $\Theta$  is compact subset of  $\mathbb{R}^d$ , the family  $\{f(\cdot|\theta), \theta \in \Theta\}$  is strongly identifiable, and for all  $x \in \mathcal{X}$ , the Hessian matrix  $D^2f(x|\theta)$  satisfies a uniform Lipschitz condition*

$$|\gamma^T(D^2f(x|\theta_1) - D^2f(x|\theta_2))\gamma| \leq C\rho(\theta_1, \theta_2)^\delta \|\gamma\|^2, \quad (3)$$

for all  $x, \theta_1, \theta_2$  and some fixed  $C$  and  $\delta > 0$ . Then, for fixed  $G_0 \in \mathcal{G}_k(\Theta)$ , where  $k < \infty$ :

$$\lim_{\epsilon \rightarrow 0} \inf_{G, G' \in \mathcal{G}_k(\Theta)} \left\{ \psi(G, G') : d_\rho(G_0, G) \vee d_\rho(G_0, G') \leq \epsilon \right\} > 0. \quad (4)$$

The assertion also holds with  $\psi$  being replaced by  $\psi_1$ .

**Remarks.** (i) In Section 5.2 the notion of strong identifiability is extended to an infinite dimensional setting via first and second order Fréchet derivatives in normed spaces.

(ii) Suppose that  $G_0$  has exactly  $k$  support points in  $\Theta$ . Then, an examination of the proof reveals that the requirement that  $\Theta$  be compact is not needed. Indeed, if there is a sequence of  $G_n \in \mathcal{G}_k(\Theta)$  such that  $d_\rho(G_0, G_n) \rightarrow 0$ , then it is simple to show that there is a subsequence of  $G_n$  that also has  $k$  distinct atoms, which converge in  $\rho$  metric to the set of  $k$  atoms of  $G_0$  (up to some permutation of the labels). The proof of the theorem proceeds as before.

For the rest of this paper, by strong identifiability we always mean conditions specified in Theorem 1 so that Eq. (4) can be deduced. This practically means that the conditions specified by Eq. (2) and Eq. (3) be given, while the compactness of  $\Theta$  may sometimes be required.

### 2.3 Wasserstein metric identifiability in infinite mixture models

Next, we state a counterpart of Theorem 1 for  $G, G' \in \bar{\mathcal{G}}(\Theta)$ , i.e., mixing measures with infinitely many support points. We restrict our attention to convolution mixture models on  $\mathbb{R}^d$ . That is, the likelihood density function  $f(x|\theta)$ , with respect to Lebesgue, takes the form  $f(x - \theta)$  for some multivariate density function  $f$  on  $\mathbb{R}^d$ . Thus,  $p_G(x) = G * f(x) = \sum_{i=1}^k p_i f(x - \theta_i)$  and  $p_{G'}(x) = G' * f(x) = \sum_{j=1}^{k'} p'_j f(x - \theta'_j)$ .

As before, we need a compactness condition for  $\Theta$ . Additional key assumptions concern with the smoothness of density function  $f$ . This is characterized in terms of the tail behavior of the Fourier transform of  $f$ . We consider both ordinary smooth densities (e.g., Laplace and Gamma), and supersmooth densities (e.g., normal). The following result does not require that  $G, G'$  be discrete.

**Theorem 2.** *Suppose that  $G, G'$  are probability measures that place full support on compact set  $\Theta \subset \mathbb{R}^d$ .  $f$  is a density function on  $\mathbb{R}^d$  that is symmetric (around 0), i.e.,  $\int_A f dx = \int_{-A} f dx$  for any Borel set  $A \subset \mathbb{R}^d$ . Moreover, assume that  $\tilde{f}(\omega) \neq 0$  for any  $\omega \in \mathbb{R}^d$ .*

- (1) **Ordinary smooth likelihood.** *If  $|\tilde{f}(\omega)| \prod_{j=1}^d |\omega_j|^\beta \geq d_0$  as  $\omega_j \rightarrow \infty$  ( $j = 1, \dots, d$ ) for some positive constant  $d_0$  and constant  $\beta$ . Then for any  $m < 4/(4 + (2\beta + 1)d)$ , there is some constant  $C(d, \beta, m)$  dependent only on  $d, \beta$  and  $m$  such that*

$$d_{\rho^2}(G, G') \leq C(d, \beta, m) d_V(p_G, p_{G'})^m,$$

as  $d_V(p_G, p_{G'}) \rightarrow 0$ .

- (2) **Supersmooth likelihood.** *If  $|\tilde{f}(\omega)| \prod_{j=1}^d \exp(|\omega_j|^\beta/\gamma) \geq d_0$  as  $\omega_j \rightarrow \infty$  ( $j = 1, \dots, d$ ) for some positive constants  $\beta, \gamma, d_0$ . There is some constant  $C(d, \beta)$  dependent only on  $d$  and  $\beta$ , such that*

$$d_{\rho^2}(G, G') \leq C(d, \beta) (-\log d_V(p_G, p_{G'}))^{-2/\beta},$$

as  $d_V(p_G, p_{G'}) \rightarrow 0$ .

**Example 2.** For the standard normal density on  $\mathbb{R}^d$ ,  $\tilde{f}(\omega) = \prod_{j=1}^d \exp -\omega_j^2/2$ , we obtain that  $d_{\rho^2}^2(G, G') \lesssim (-\log d_V(p_G, p_{G'}))^{-1}$  as  $d_{\rho}(G, G') \rightarrow 0$  (so that  $d_V(p_G, p_{G'}) \rightarrow 0$ , by Lemma 1). For a Laplace density on  $\mathbb{R}$ , e.g.,  $\tilde{f}(\omega) = \frac{1}{1+\omega^2}$ , then  $d_{\rho^2}^2(G, G') \lesssim d_V(p_G, p_{G'})^m$  for any  $m < 4/(4 + 5d)$ , as  $d_{\rho}(G, G') \rightarrow 0$ .

### 3 Convergence of posterior distributions of mixing measures

We are ready to study the convergence of discrete mixing measures in a Bayesian setting. Let  $X_1, \dots, X_n$  be an iid sample according to the mixture density  $p_G(x) = \int f(x|\theta) dG(\theta)$ , where  $f$  is known, while  $G = G_0$  for some unknown mixing measure in  $\mathcal{G}_k(\Theta)$ . The number support points  $k$  may not be known. In the Bayesian framework,  $G$  is endowed with a prior distribution  $\Pi$  on a suitable measure space of discrete probability measures in  $\bar{\mathcal{G}}(\Theta)$ . The posterior distribution of  $G$  is given by, for any measurable set  $B$ :

$$\Pi(B|X_1, \dots, X_n) = \int_B \prod_{i=1}^n p_G(X_i) \Pi(G) / \int \prod_{i=1}^n p_G(X_i) \Pi(G).$$



We shall study conditions under which the posterior distribution is consistent, i.e., it concentrates on arbitrarily small  $d_\rho$  neighborhoods of  $G_0$ , and establish the rates of the convergence. The analysis is based upon the general framework of Ghosal, Ghosh and van der Vaart Ghosal et al. (2000), who analyzed convergence behavior of posterior distributions in terms of  $f$ -divergences such as Hellinger and variational distances on the mixture densities of the data. In the following we formulate two convergence theorems for mixture model setting (which can be viewed as counterparts of Theorem 2.1 and 2.4 of Ghosal et al. (2000)). A notable feature of our theorems is that conditions (e.g., entropy and prior concentration) are stated in terms of the Wasserstein metric, as opposed to  $f$ -divergences on the mixture densities. They may be typically separated into independent conditions for the prior for  $G$  and the likelihood family and are simpler to verify for mixture models. In addition, the convergence of the posterior distribution of mixing measures is established in terms of Wasserstein distance metrics.

The following notion plays a central role in the theorem's formulation.

**Definition 6.** Let  $\mathcal{G}$  be a subset of  $\bar{\mathcal{G}}(\Theta)$ . For each  $k < \infty$ , define the Hellinger information of  $d_\rho$  metric as a real-valued function on the real line  $C_k(\mathcal{G}, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$ :

$$C_k(\mathcal{G}, r) = \inf_{G_0 \in \mathcal{G}_k(\Theta), G \in \mathcal{G} : d_\rho(G_0, G) \geq r/2} d_h^2(p_{G_0}, p_G)/2. \quad (5)$$

It is obvious that  $C_k$  is a non-negative and non-decreasing function. The following characterization of  $C_k$  follows immediately from the results obtained in the previous section.

**Proposition 1.** (a) If  $\mathcal{G}$  and  $\mathcal{G}_k(\Theta)$  are both compact in the Wasserstein topology, and the family of likelihood functions is finitely identifiable. Then,  $C_k(\mathcal{G}, r) > 0$  for any  $r > 0$ .

(b) If  $\Theta \subset \mathbb{R}^d$  is compact, and the family of likelihood functions is strongly identifiable as specified in Theorem 1. Then, for each  $k$  there is a constant  $c(k) > 0$  such that  $C_k(\mathcal{G}_k(\Theta), r) \geq c(k)r^4$  for all  $r > 0$ .

(c) If  $\Theta \in \mathbb{R}^d$  is compact, and the family of likelihood functions is ordinary smooth with parameter  $\beta$ , as specified in Theorem 2. Then, for any  $\epsilon > 0$ , there is some constant  $c(d, \beta)$  such that  $C_k(\bar{\mathcal{G}}(\Theta), r) \geq c(d, \beta)r^{4+(2\beta+1)d+\epsilon}$  for any  $r > 0$ , and any  $k > 0$ . For supersmooth likelihood family, we have  $C_k(\bar{\mathcal{G}}(\Theta), r) \geq \exp[-c(d, \beta)r^{-\beta}]$  for any  $r > 0$  and any  $k > 0$ .

The following two theorems have three types of conditions. The first is concerned with the size of support of  $\Pi$ , often quantified in terms of its entropy number. Estimates of the entropy number defined in terms of Wasserstein metrics for several measure classes of interest are given in Lemma 2. The second is on the Kullback-Leibler support of  $\Pi$ , which is related to both space of discrete measures  $\bar{\mathcal{G}}(\Theta)$  and the family of likelihood functions  $f(x|\theta)$ . The Kullback-Leibler neighborhood is defined as:

$$B(\epsilon) = \left\{ G \in \bar{\mathcal{G}}(\Theta) : -P_{G_0} \left( \log \frac{p_G}{p_{G_0}} \right) \leq \epsilon^2, P_{G_0} \left( \log \frac{p_G}{p_{G_0}} \right)^2 \leq \epsilon^2 \right\}. \quad (6)$$

The third condition is on the Hellinger information of  $d_\rho$  metric, function  $C_k(\mathcal{G}, r)$ , a characterization of which is given above.

**Theorem 3.** *Let  $G_0 \in \mathcal{G}_k(\Theta) \subseteq \bar{\mathcal{G}}(\Theta)$  for some  $k < \infty$ , and the family of likelihood functions is finitely identifiable. Suppose that for a sequence  $(\epsilon_n)_{n \geq 1}$  that tends to a constant (or 0) such that  $n\epsilon_n^2 \rightarrow \infty$ , sets  $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$  and a constant  $C > 0$ , we have*

$$\log D(\epsilon_n, \mathcal{G}_n, d_\rho) \leq n\epsilon_n^2, \quad (7)$$

$$\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) \leq \exp[-n\epsilon_n^2(C + 4)], \quad (8)$$

$$\Pi(B(\epsilon_n)) \geq \exp(-n\epsilon_n^2 C). \quad (9)$$

Moreover, suppose  $M_n$  is a sequence so that

$$C_k(\mathcal{G}_n, M_n \epsilon_n) \geq \epsilon_n^2(C + 4), \quad (10)$$

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)] \rightarrow 0. \quad (11)$$

Then,  $\Pi(G : d_\rho(G_0, G) \geq M_n \epsilon_n | X_1, \dots, X_n) \rightarrow 0$  in  $P_{G_0}$ -probability.

A stronger theorem (using a substantially weaker condition on the covering number) can be formulated as follows.

**Theorem 4.** *Let  $G_0 \in \mathcal{G}_k(\Theta) \subseteq \bar{\mathcal{G}}(\Theta)$  for some  $k < \infty$ , and the family of likelihood functions is finitely identifiable. Suppose that for a sequence  $\epsilon_n \rightarrow 0$  such that  $n\epsilon_n^2$  is bounded away from 0 or tending to infinity, and sets  $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$ , we have*

$$\log D(\epsilon/2, \{G \in \mathcal{G}_n : \epsilon \leq d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq n\epsilon_n^2 \text{ for every } \epsilon \geq \epsilon_n, \quad (12)$$

$$\frac{\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n)}{\Pi(B(\epsilon_n))} = o(\exp(-2n\epsilon_n^2)), \quad (13)$$

$$\frac{\Pi(G : j\epsilon_n < d_\rho(G, G_0) \leq 2j\epsilon_n)}{\Pi(B(\epsilon_n))} \leq \exp[nC_k(\mathcal{G}_n, j\epsilon_n)/2] \text{ for any } j \geq M_n, \quad (14)$$

where  $M_n$  is a sequence such that

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/2] \rightarrow 0. \quad (15)$$

Then, we have that  $\Pi(G : d_\rho(G_0, G) \geq M_n \epsilon_n | X_1, \dots, X_n) \rightarrow 0$  in  $P_{G_0}$ -probability.

**Remarks.** (i) The above statement continues to hold if conditions (14) and (15) are replaced by the following condition:

$$\exp(2n\epsilon_n^2)/\Pi(B(\epsilon_n)) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)] \rightarrow 0. \quad (16)$$

(ii) The above theorems are stated for the  $L_1$  Wasserstein metric  $d_\rho$ , but they also hold for  $L_2$  Wasserstein metric  $d_{\rho^2}^{1/2}$ , with a slight modification of the definition of the Hellinger information function to  $C_k(\mathcal{G}, r) = \inf_{d_{\rho^2}^{1/2}(G_0, G) \geq r/2} d_h^2(p_{G_0}, p_G)$ .

Before moving to specific examples, we state a simple lemma, which provides estimates of the entropy under  $d_\rho$  metric for a number of classes of discrete measures of interest. Because  $d_\rho$  inherits directly the  $\rho$  metric in  $\Theta$ , the entropy for classes in  $(\bar{\mathcal{G}}(\Theta), d_\rho)$  can typically be bounded in terms of the covering number for subsets of  $(\Theta, \rho)$ . These bounds will be used extensively in the sequel.

**Lemma 2.** (a)  $\log N(2\epsilon, \mathcal{G}_k(\Theta), d_\rho) \leq k(\log N(\epsilon, \Theta, \rho) + \log(e + e\text{Diam}(\Theta)/\epsilon))$ .

(b)  $\log N(2\epsilon, \bar{\mathcal{G}}(\Theta), d_\rho) \leq N(\epsilon, \Theta, \rho) \log(e + e\text{Diam}(\Theta)/\epsilon)$ .

(c) Let  $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$ . Assume that  $M = \max_{i=1}^k 1/p_i^* < \infty$  and  $m = \min_{i,j \leq k} \rho(\theta_i^*, \theta_j^*) > 0$ . Then,

$$\begin{aligned} \log N(\epsilon/2, \{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \\ \leq k(\sup_{\Theta'} \log N(\epsilon/4, \Theta', \rho) + \log(32k\text{Diam}(\Theta)/m)), \end{aligned}$$

where the supremum in the right side is taken over all bounded subsets  $\Theta' \subseteq \Theta$  such that  $\text{Diam}(\Theta') \leq 4M\epsilon$ .

## 4 Examples

In this section the general theory is illustrated in specific mixture models, including finite mixtures of multivariate distributions, infinite mixtures based on Dirichlet processes, and finite mixtures of Gaussian processes.

### 4.1 Finite mixture of multivariate distributions

Let  $\Theta$  be a subset of  $\mathbb{R}^d$ ,  $\rho$  be the Euclidean metric, and  $\Pi$  is a prior distribution for discrete measures in  $\mathcal{G}_k(\Theta)$ , where  $k < \infty$  is known. Suppose that the “truth”  $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$ . To obtain the convergence rate of the posterior distribution of  $G$ , we need:

#### Assumptions A.

- (A1)  $\Theta$  is compact and the family of likelihood functions  $f(\cdot|\theta)$  is strongly identifiable.
- (A2) For some positive constants  $C_1, C_2$ ,  $d_K(f_i, f'_j) \leq C_1 \rho^2(\theta_i, \theta'_j)$  and  $\int f_i [\log(f_i/f'_j)]^2 \leq C_2 \rho^2(\theta_i, \theta'_j)$  for any  $\theta_i, \theta'_j \in \Theta$ .
- (A3) Under prior  $\Pi$ , for small  $\delta > 0$ ,  $c_3 \delta^k \leq \Pi(|p_i - p_i^*| \leq \delta, i = 1 \dots, k) \leq C_3 \delta^k$  and  $c_3 \delta^{kd} \leq \Pi(\rho(\theta_i, \theta_i^*) \leq \delta, i = 1 \dots, k) \leq C_3 \delta^{kd}$  for some constants  $c_3, C_3 > 0$ .

**Remarks.** (A1) and (A2) hold for the family of Gaussian densities with mean parameter  $\theta$ . (A3) holds when the prior distribution on the relevant parameters behave like a uniform distribution, up to a multiplicative constant.

Let  $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ . Combining Lemma 1 with Assumption (A2), if  $\rho(\theta_i, \theta_i^*) \leq \epsilon$  and  $|p_i - p_i^*| \leq \epsilon^2/(k \text{Diam}(\Theta)^2)$  for  $i = 1, \dots, k$ , then  $d_K(p_{G_0}, p_G) \leq d_{\rho K}(G_0, G) \leq C_1 \sum_{1 \leq i, j \leq k} q_{ij} \rho^2(\theta_i^*, \theta_j)$ , for any  $\mathbf{q} \in \mathcal{Q}$ . Thus,  $d_K(p_{G_0}, p_G) \leq C_1 d_{\rho^2}(G_0, G) \leq C_1 \sum_{i=1}^k (p_i^* \wedge p_i) \rho^2(\theta_i^*, \theta_i) + C_1 \sum_{i=1}^k |p_i - p_i^*| \text{Diam}(\Theta)^2 \leq 2C_1 \epsilon^2$ . Hence, under prior  $\Pi$ ,

$$\Pi(G : d_K(p_{G_0}, p_G) \leq \epsilon^2) \geq \Pi(G : \rho(\theta_i, \theta_i^*) \leq \epsilon, |p_i - p_i^*| \leq \epsilon^2/(k \text{Diam}(\Theta)^2), i = 1, \dots, k).$$

Similarly, due to (A2),  $\int p_{G_0} [\log(p_{G_0}/p_G)]^2 \leq C_2 d_{\rho^2}(G_0, G)$ . Thus, in view of Assumption (A3), we have  $\Pi(B(\epsilon)) \gtrsim \epsilon^{k(d+2)}$ . Conversely, for sufficiently small  $\epsilon$ , if  $d_{\rho^2}(G_0, G) \leq \epsilon^2$  then by reordering the index of the atoms, we must have  $\rho(\theta_i, \theta_i^*) = O(\epsilon)$  and  $|p_i - p_i^*| = O(\epsilon^2)$  for all  $i = 1, \dots, k$  (see the argument in the proof of Lemma 2(c)). This entails that under the prior  $\Pi$ ,

$$\Pi(G : d_{\rho^2}(G_0, G) \leq \epsilon^2) \leq \Pi(G : \rho(\theta_i, \theta_i^*) \leq O(\epsilon), |p_i - p_i^*| \leq O(\epsilon^2), i = 1, \dots, k) \lesssim \epsilon^{k(d+2)}.$$

Let  $\epsilon_n = n^{-1/2}$ . We proceed by verifying conditions of Theorem 4, as this theorem provides the right rate for parametric mixture models under the  $L_2$  Wasserstein distance metric  $d_{\rho^2}^{1/2}$ . Let  $\mathcal{G}_n := \mathcal{G}_k(\Theta)$ . Then  $\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) = 0$ , so Eq. (13) trivially holds.

Next, we show that  $D(\epsilon/2, S, d_{\rho^2}^{1/2})$ , where  $S = \{G \in \mathcal{G}_n : d_{\rho}(G_0, G) \leq 2\epsilon\}$ , is bounded above by a constant, so that (12) is satisfied. Indeed, for any  $\epsilon > 0$ ,  $\log D(\epsilon/2, S, d_{\rho^2}^{1/2}) \leq \log N(\epsilon/4, S, d_{\rho^2}^{1/2}) \leq N(\epsilon/4, S, d_{\rho})$ . By Lemma 2 (c),  $N(\epsilon/4, S, d_{\rho})$  is bounded in terms of  $\sup_{\Theta'} \log N(\epsilon/8, \Theta', \rho)$ , which is bounded above by a constant when  $\Theta'$ 's are subsets of  $\Theta$  whose diameter is bounded by a multiple of  $\epsilon$ . Thus, Eq. (12) holds.

By Proposition 1(b) and Assumption (A4), there is  $C_k(\mathcal{G}_n, j\epsilon_n) = \inf_{d_{\rho^2}(G_0, G) \geq (j\epsilon/2)^2} d_h^2(p_{G_0}, p_G) \geq c(j\epsilon_n)^4$  for some constant  $c > 0$ . To ensure condition (15), note that:

$$\begin{aligned} \exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/2] &\leq \exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nc(j\epsilon_n)^4] \\ &\lesssim \exp(2n\epsilon_n^2 - ncM_n^4\epsilon_n^4). \end{aligned}$$

This upper bound goes to zero if  $ncM_n^4\epsilon_n^4 \geq 4n\epsilon_n^2$ , which is satisfied by taking  $M_n$  to be a large multiple of  $\epsilon_n^{-1/2}$ . Thus we need  $M_n\epsilon_n \asymp \epsilon_n^{1/2} = n^{-1/4}$ .

Under the assumptions specified above,  $\Pi(G : j\epsilon_n < d_{\rho}(G, G_0) \leq 2j\epsilon_n)/\Pi(B(\epsilon_n)) = O(1)$ . On the other hand, for  $j \geq M_n$ , we have  $\exp[nC_k(\mathcal{G}_n, j\epsilon_n)/2] \geq \exp[nc(j\epsilon_n)^4/2]$  which is bounded below by arbitrarily large constant by choosing  $M_n$  to be a large multiple of  $\epsilon_n^{-1/2}$ , thereby ensuring (14).

Thus, by Theorem 4, rate of contraction for the posterior distribution of  $G$  under  $d_{\rho^2}^{1/2}$  distance metric is  $n^{-1/4}$ , which is also the minimax rate  $n^{-1/4}$  as proved in the univariate case by Chen (1995). Our calculation is summarized by:

**Theorem 5.** *Under Assumptions (A1–A3), the contraction rate in the  $L_2$  Wasserstein distance metric of the posterior distribution of  $G$  is  $n^{-1/4}$ .*

## 4.2 Infinite mixture based on the Dirichlet process

Given the “true” discrete measure  $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$ , where  $\Theta$  is a metric space but  $k \leq \infty$  is unknown. To estimate  $G_0$ , the prior distribution  $\Pi$  on discrete measure  $G \in \bar{\mathcal{G}}(\Theta)$  is taken to be a Dirichlet process  $\text{DP}(\nu, P_0)$  that centers at  $P_0$  with concentration parameter  $\nu > 0$  Ferguson (1973). Here, parameter  $P_0$  is a probability measure on  $\Theta$ . For any  $m \geq 1$ , the following lemma provides a lower bound of small ball probabilities of metric space  $(\bar{\mathcal{G}}(\Theta), d_{\rho^m}^{1/m})$  in terms of small probabilities of metric space  $(\Theta, \rho)$ .

**Lemma 3.** *Let  $G \sim \text{DP}(\nu, P_0)$ , where  $P_0$  is a non-atomic base probability measure on a compact set  $\Theta$ . For a small  $\epsilon > 0$ , let  $D = D(\epsilon, \Theta, \rho)$  denote the packing number of  $\Theta$  under  $\rho$  metric. Then, under the Dirichlet process distribution,*

$$\Pi(G : d_{\rho^m}(G_0, G) \leq (2^m + 1)\epsilon^m) \geq \Gamma(\nu)(\epsilon^m D^{-1} / \text{Diam}(\Theta)^m)^{D-1} \nu^D \prod_{i=1}^D P_0(S_i).$$

Here,  $(S_1, \dots, S_D)$  denotes the  $D$  disjoint  $\epsilon/2$ -balls that form a maximal packing of  $\Theta$ .  $\Gamma(\cdot)$  is the gamma function.

*Proof.* Since every point in  $\Theta$  is of distance at most  $\epsilon$  to one of the centers of  $S_1, \dots, S_D$ , there is a  $D$ -partition  $(S'_1, \dots, S'_D)$  of  $\Theta$ , such that  $S_i \subseteq S'_i$ , and  $\text{Diam}(S'_i) \leq 2\epsilon$ . for each  $i = 1, \dots, D$ . Let  $m_i = G(S'_i)$ ,  $\mu_i = P_0(S'_i)$ , and  $\hat{p}_i = G_0(S'_i)$ . From the definition of Dirichlet processes,  $\mathbf{m} = (m_1, \dots, m_D) \sim \text{Dir}(\nu\mu_1, \dots, \nu\mu_D)$ . Note that

$$d_{\rho^m}(G_0, G) \leq (2\epsilon)^m + \|\mathbf{m} - \hat{\mathbf{p}}\|_1 [\text{Diam}(\Theta)]^m.$$

Due to the non-atomicity of  $P_0$ , for  $\epsilon$  sufficiently small,  $\nu\mu_i \leq 1$  for all  $i = 1, \dots, D$ . Let  $\delta = \epsilon / \text{Diam}(\Theta)$ . Then, under  $\Pi$ ,

$$\begin{aligned} \Pr(d_{\rho^m}(G_0, G) \leq (2^m + 1)\epsilon^m) &\geq \Pr(\|\mathbf{m} - \hat{\mathbf{p}}\|_1 \leq \delta^m) \geq \Pr(|m_i - \hat{p}_i| \leq \delta^m / D, i = 1, \dots, D) \\ &= \frac{\Gamma(\nu)}{\prod_{i=1}^D \Gamma(\nu\mu_i)} \int_{\Delta_{D-1} \cap \{|m_i - \hat{p}_i| \leq \delta^m / D\}} \prod_{i=1}^{D-1} m_i^{\nu\mu_i-1} (1 - \sum_{i=1}^{D-1} m_i)^{\nu\mu_D-1} dm_1 \dots dm_{D-1} \\ &\geq \frac{\Gamma(\nu)}{\prod_{i=1}^D \Gamma(\nu\mu_i)} \prod_{i=1}^{D-1} \int_{\max(\hat{p}_i - \delta^m / D, 0)}^{\min(\hat{p}_i + \delta^m / D, 1)} m_i^{\nu\mu_i-1} dm_i \geq \Gamma(\nu)(\delta^m / D)^{D-1} \prod_{i=1}^D (\nu\mu_i). \end{aligned}$$

The second inequality is due to  $(1 - \sum_{i=1}^{D-1} m_i)^{\nu\mu_D-1} = m_D^{\nu\mu_D-1} \geq 1$ , since  $\nu\mu_D \leq 1$  and  $0 < m_D < 1$  almost surely. The third inequality is due to the fact that  $\Gamma(\alpha) \leq 1/\alpha$  for  $0 < \alpha \leq 1$ . This gives the desired claim.  $\square$

### Assumptions C.

- (C1) The non-atomic base measure  $P_0$  places full support on a compact set  $\Theta$ . The family of the likelihood densities  $f(\cdot|\theta)$  is finitely identifiable.
- (C2) For some constants  $C_1, m_1, C_2, m_2 > 0$ ,  $d_K(f_i, f_j') \leq C_1 \rho^{m_1}(\theta_i, \theta_j')$  and  $\int f_i [\log(f_i/f_j')]^2 \leq C_1 \rho^{m_2}(\theta_i, \theta_j')$  for any  $\theta_i, \theta_j' \in \Theta$ .
- (C3)  $P_0$  places sufficient probability mass on all small balls that pack  $\Theta$ . Specifically, there is a universal constant  $c_3 > 0$  such that the probability of the  $D$ -partition  $(S_1, \dots, S_D)$  specified in Lemma 3 satisfy for any  $\epsilon > 0$ :

$$\log \prod_{i=1}^D P_0(S_i) \geq c_3 D \log(1/D).$$

- (C4)  $\Theta \subset \mathbb{R}^d$  is compact, so that the packing number  $D(\epsilon, \Theta, \rho) \asymp [\text{Diam}(\Theta)/\epsilon]^d$ .

**Theorem 6.** *Given Assumptions (C1–C4), and the smoothness conditions for the likelihood family as specified in Theorem 2, there is a sequence  $\beta_n \searrow 0$  such that  $\Pi(d_\rho(G_0, G) \geq \beta_n | X_1, \dots, X_n) \rightarrow 0$  in  $P_{G_0}$  probability. Specifically,*

- (1) *For ordinary smooth likelihood functions, take  $\beta_n \asymp (\log n/n)^{\frac{2}{(d+2)(4+(2\beta+1)d)+\delta}}$ , for any small  $\delta > 0$ .*
- (2) *For supersmooth likelihood functions, take  $\beta_n \asymp (\log n)^{-1/\beta}$ .*

*Proof.* The proof consists of two main steps. First, we shall prove that under Assumptions (C1–C4), conditions specified by Eqs. (7) (8) (9) in Theorem 3 are satisfied by taking  $\mathcal{G}_n = \bar{\mathcal{G}}(\Theta)$ , and  $\epsilon_n$  to be a large multiple of  $(\log n/n)^{1/(d+2)}$ . The second step involves constructing a sequence of  $M_n$  and subsequently  $\beta_n = M_n \epsilon_n$  for which Theorem 3 can be applied.

Step 1: By Lemma 1 and (C2),  $d_K(p_{G_0}, p_G) \leq d_{\rho K}(G_0, G) \leq C_1 \rho^{m_1}(G_0, G)$ . Also,  $\int p_{G_0} [\log(p_{G_0}/p_G)]^2 \leq C_2 \rho^{m_2}(G_0, G)$ . Without loss of generality, assume that  $m_1 \leq m_2$ . We obtain that  $\Pi(G \in B(\epsilon_n)) \geq \Pi(G : d_{\rho^{m_1}}(G_0, G) \leq C_3 \epsilon_n^2)$  for some constant  $C_3$ . Combining this bound with (C3) and (C4), which are applied to Lemma 3 we have:  $\log \Pi(G \in B(\epsilon_n)) \gtrsim (D-1) \log(\epsilon_n/\text{Diam}(\Theta)) + (2D-1) \log(1/D) + D \log \nu$ , where the approximation constant is dependent on  $m_1, m_2$ . Note that  $D \asymp [\text{Diam}(\Theta)/\epsilon_n]^d$ . It is simple to check that condition (9) holds,  $\log \Pi(G \in B(\epsilon_n)) \geq -C n \epsilon_n^2$ , by the given rate of  $\epsilon_n$ , for any constant  $C > 0$ .

Since  $\mathcal{G}_n = \bar{\mathcal{G}}(\Theta)$ , (8) trivially holds. Turning to condition (7), by Lemma 2(b), we have  $\log N(2\epsilon_n, \bar{\mathcal{G}}(\Theta), d_\rho) \leq N(\epsilon_n, \Theta, \rho) \log(e + e \text{Diam}(\Theta)/\epsilon_n) \leq (\text{Diam}(\Theta)/\epsilon_n)^d \log(e + e \text{Diam}(\Theta)/\epsilon_n) \leq n \epsilon_n^2$  by the specified rate of  $\epsilon_n$ .

Step 2: For any  $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$ , let  $R_k(\mathcal{G}, r)$  be the inverse of the Hellinger information function of  $d_\rho$  metric. Specifically, for any  $t \geq 0$ ,

$$R_k(\mathcal{G}, t) = \inf\{r \geq 0 | C_k(\mathcal{G}, r) \geq t\}.$$

Note that  $R_k(\mathcal{G}, 0) = 0$ .  $R_k(\mathcal{G}, \cdot)$  is non-decreasing because  $C_k(\mathcal{G}, \cdot)$  is.

Let  $(\epsilon_n)_{n \geq 1}$  be the sequence determined in the previous step of the proof. Let  $M_n = R_k(\bar{\mathcal{G}}(\Theta), \epsilon_n^2(C+4))/\epsilon_n$ , and  $\beta_n = M_n \epsilon_n = R_k(\bar{\mathcal{G}}(\Theta), \epsilon_n^2(C+4))$ . Condition (10) holds by definition of  $R_k$ , i.e.,  $C_k(\mathcal{G}(\Theta), M_n \epsilon_n) \geq \epsilon_n^2(C+4)$ . To verify (11), note that the running sum with respect to  $j$  cannot have more than  $\text{Diam}(\Theta)/\epsilon_n$ , and due to the monotonicity of  $C_k$ , we have

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)] \leq \text{Diam}(\Theta)/\epsilon_n \exp(2n\epsilon_n^2 - nC_k(\mathcal{G}_n, M_n\epsilon_n)) \rightarrow 0.$$

Hence, Theorem 3 can be applied to conclude that  $\Pi(d_\rho(G_0, G) \geq \beta_n | X_1, \dots, X_n) \rightarrow 0$  in  $P_{G_0}$  probability. Under ordinary smoothness condition (as specified in Theorem 2),

$$R_k(\bar{\mathcal{G}}(\Theta), t) = t^{\frac{1}{4+(2\beta+1)d+\delta}}, \text{ where } \delta \text{ is an arbitrarily positive constant. So, } \beta_n \asymp \epsilon_n^{\frac{2}{4+(2\beta+1)d+\delta}} = (\log n/n)^{\frac{2}{(d+2)(4+(2\beta+1)d+\delta)}}. \text{ On the other hand, under supersmoothness condition, } R_k(\bar{\mathcal{G}}(\Theta), t) = (1/\log(1/t))^{1/\beta}. \text{ So, } \beta_n \asymp (\log(1/\epsilon_n))^{-1/\beta} \asymp (\log n)^{-1/\beta}.$$

□

### 4.3 Finite mixture of Gaussian processes

We now study an example in which  $\Theta$  has infinite dimensions. Specifically, let  $T = [0, 1]$ , and  $\Theta = l_\infty(T)$  is the Banach space of bounded functions  $\theta : T \rightarrow \mathbb{R}$ , equipped with the uniform norm  $\|\theta\| = \sup\{|\theta(t)| : t \in T\}$ . Suppose that the “true”  $G_0$  has  $k$  distinct atoms in  $\Theta$ ,  $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*}$ , where  $k$  is known.

We shall consider a “mixture of Gaussian processes” prior  $\Pi$  on  $\mathcal{G}_k(\Theta)$ . Specifically, a random draw  $G$  from  $\Pi$  is a discrete measure taking the form  $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ , where  $\theta_i$  are independent random sample paths distributed according to a zero-mean Gaussian process. Let  $K : T \times T \rightarrow \mathbb{R}$  be the covariance function that defines the Gaussian process — we assume further that the Gaussian process has bounded sample paths and that it is Borel measurable.

It is a known fact that the support of the defined zero-mean Gaussian measure is equal to the closure of the reproducing kernel Hilbert space (RKHS), to be denoted by  $\mathbb{H}(K)$ , or simply  $\mathbb{H}$ , of the covariance kernel  $K$  of the process (Kallianpur, 1971; van der Vaart and van Zanten, 2008a). Assume that  $\theta_i^* \in \bar{\mathbb{H}}$  for all  $i = 1, \dots, k$ , so that the “true”  $G_0$  is contained within the support of prior  $\Pi$ .

An ingredient of our analysis is drawn from a recent result by van der Vaart and van Zanten, who studied asymptotic behavior of priors based on Gaussian processes van der Vaart and van Zanten (2008b). Define  $\rho(\theta_i, \theta_j) = \sup |\theta_i(t) - \theta_j(t)| : t \in T$ . A key notion in their analysis is concentration functions. For each  $i = 1, \dots, k$ , define the concentration function:

$$\phi_{\theta_i^*}(\epsilon) = \inf_{h \in \mathbb{H} : \rho(h, \theta_i^*) < \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|\theta\| < \epsilon),$$

where  $\Pr$  denotes the probability under the Gaussian process.

Another ingredient is the extension of the notion of strong identifiability to function space  $\Theta$ , see Section 5.2, so that the result of Theorem 1 continues to hold for the infinite dimensional  $\Theta$ . Finally, we need additional assumptions:

### Assumptions B.

- (B1) The family of likelihood functions  $\{f(\cdot|\theta), \theta \in \Theta\}$  is strongly identifiable.
- (B2) For some positive constants  $C_1, C_2$ ,  $d_K(f_i, f'_j) \leq C_1 \rho^2(\theta_i, \theta'_j)$  and  $\int f_i [\log(f_i/f'_j)]^2 \leq C_2 \rho^2(\theta_i, \theta'_j)$  for any  $\theta_i, \theta'_j \in \Theta$ .
- (B3) Under prior  $\Pi$ , for small  $\delta > 0$ ,  $\Pi(|p_i - p_i^*| \leq \delta, i = 1 \dots, k) \geq c_3 \delta^{k\alpha}$  for some constants  $c_3, \alpha > 0$ .
- (B4) Under prior  $\Pi$ ,  $C_4 = \mathbb{E}\|\theta\|^2 < \infty$ .

**Theorem 7.** *Given Assumptions (B1–B4). Let  $(\epsilon_n)_{n \in \mathbb{N}}$  be a sequence of positive numbers tending to 0 such that  $\log n = o(n\epsilon_n^2)$  and that for any  $i = 1, \dots, k$  and any  $n$ ,*

$$\phi_{\theta_i^*}(\epsilon_n) \leq n\epsilon_n^2. \quad (17)$$

*Then, for a sufficiently large constant  $M > 0$ ,  $\Pi(d_\rho(G_0, G) \geq M\epsilon_n^{1/2} | X_1, \dots, X_n) \rightarrow 0$  in  $P_{G_0}$ -probability.*

**Remarks.** (i) The reader is referred to van der Vaart and van Zanten (2008b) for examples of convergence rates  $\epsilon_n$  that satisfy condition (17) for different choices of  $\Theta$ . In particular, if under the Gaussian process prior, the supporting atoms  $\theta_i$  of  $G \sim \Pi$  are functions on  $T = [0, 1]$  with smoothness  $\gamma_1 > 0$ , while the “true” support points  $\theta_i^*$ ’s of  $G_0$  are functions with smoothness  $\gamma_2 > 0$ , then concentration functions  $\phi_{\theta_i^*}(\epsilon) = \epsilon^{-1/(\gamma_1 \wedge \gamma_2)}$  for each  $i = 1, \dots, k$ . Accordingly, the rate  $\epsilon_n$  for which Eq. (17) holds is  $\epsilon_n \asymp n^{-\frac{\gamma_1 \wedge \gamma_2}{2\gamma_1 \wedge \gamma_2 + 1}}$ . The contraction rate for the posterior distribution of  $G$  is  $n^{-\frac{\gamma_1 \wedge \gamma_2}{2(2\gamma_1 \wedge \gamma_2 + 1)}}$ .

(ii) We currently do not have a concrete example of likelihood functions  $f(\cdot|\theta)$  satisfying strong identifiability conditions. We plan to explore this issue in a future work.

## 5 Proofs of main results

### 5.1 Identifiability results

#### Proof of Theorem 1.

*Proof.* Suppose that Eq. (4) is not true, then there will be sequences of  $G_n$  and  $G'_n$  tending to  $G_0$  in  $d_\rho$  metric, and that  $\psi(G_n, G'_n) \rightarrow 0$ . We write  $G_n = \sum_{i=1}^\infty p_{n,i} \delta_{\theta_{n,i}}$ , where  $p_{n,i} = 0$  for indices  $i$  greater than  $k_n$ , the number of atoms of  $G_n$ . Similar notation is applied to  $G'_n$ . Since both  $G_n$  and  $G'_n$  have finite number of atoms, there is  $\mathbf{q}^{(n)} \in \mathcal{Q}(\mathbf{p}_n, \mathbf{p}'_n)$



so that  $d_{\rho^2}(G_n, G'_n) = \sum_{ij} q_{ij}^{(n)} \rho^2(\theta_{n,i}, \theta'_{n,j})$ . Note that  $d_{\rho^2}^2(G_n, G'_n) \leq d_{\rho^2}(G_n, G'_n) = O(d_{\rho}(G_n, G'_n))$ , while the latter inequality is due to the boundedness of  $\Theta$ .

Let  $\mathcal{O}_n = \{(i, j) : \rho^2(\theta_{n,i}, \theta'_{n,j}) \leq d_{\rho^2}^{1-\delta'}(G_n, G'_n)\}$  for some  $\delta' \in (0, 1)$ . Then,  $\sum_{(i,j) \notin \mathcal{O}_n} q_{ij}^{(n)} \leq d_{\rho^2}(G_n, G'_n) / d_{\rho^2}^{1-\delta'}(G_n, G'_n) \rightarrow 0$  as  $n \rightarrow \infty$ . Since  $\mathbf{q}^{(n)} \in \mathcal{Q}(\mathbf{p}_n, \mathbf{p}'_n)$ , we can express

$$\begin{aligned} \psi(G_n, G'_n) &= \sup_x \left| \sum_{i=1}^{k_n} p_{n,i} f(x|\theta_{n,i}) - \sum_{j=1}^{k'_n} p'_{n,j} f(x|\theta'_{n,j}) \right| / d_{\rho^2}(G_n, G'_n) \\ &= \sup_x \left| \sum_{ij} q_{ij}^{(n)} (f(x|\theta_{n,i}) - f(x|\theta'_{n,j})) \right| / d_{\rho^2}(G_n, G'_n), \end{aligned}$$

and, by Taylor's expansion,

$$\begin{aligned} \psi(G_n, G'_n) &= \sup_x \left| \sum_{(i,j) \notin \mathcal{O}_n} q_{ij}^{(n)} (f(x|\theta'_{n,j}) - f(x|\theta_{n,i})) + \right. \\ &\quad \sum_{(i,j) \in \mathcal{O}_n} q_{ij}^{(n)} (\theta'_{n,j} - \theta_{n,i})^T Df(x|\theta_{n,i}) \\ &\quad \sum_{(i,j) \in \mathcal{O}_n} q_{ij}^{(n)} (\theta'_{n,j} - \theta_{n,i})^T D^2 f(x|\theta_{n,i}) (\theta'_{n,j} - \theta_{n,i}) + \\ &\quad \left. R_n(x) \right| / d_{\rho^2}(G_n, G'_n) \\ &= \sup_x |A_n(x) + B_n(x) + C_n(x) + R_n(x)| / D_n, \end{aligned}$$

where

$$R_n(x) = O\left( \sum_{(i,j) \in \mathcal{O}_n} q_{ij}^{(n)} \rho^{2+\delta}(\theta_{n,i}, \theta'_{n,j}) \right) = O\left( \sum_{(i,j) \in \mathcal{O}_n} q_{ij}^{(n)} \rho^2(\theta_{n,i}, \theta'_{n,j}) d_{\rho^2}^{(1-\delta')\delta/2}(G_n, G'_n) \right)$$

due to Eq. (3) and the definition of  $\mathcal{O}_n$ . So  $R_n(x)/d_{\rho^2}(G_n, G'_n) \rightarrow 0$ . The quantities  $A_n(x)$ ,  $B_n(x)$  and  $C_n(x)$  are linear functionals of  $f(x|\theta)$ ,  $Df(x|\theta)$  and  $D^2 f(x|\theta)$  for different  $\theta$ 's, respectively. Since  $\Theta$  is compact, subsequences of  $G_n$  and  $G'_n$  can be chosen so that each of their support points converges to a fixed atom  $\theta_l^*$ , for  $l = 1, \dots, k^* \leq k$ .

After being properly rescaled, the limits of  $A_n(x)$ ,  $B_n(x)$  and  $C_n(x)$  are still linear functionals with constant coefficients not depending on  $x$ . In particular,  $C_n(x)/D_n \rightarrow \sum_{j=1}^{k^*} \gamma_j^T D^2 f(x|\theta_j^*) \gamma_j$  for some  $\gamma_j$ 's and not all these coefficients vanishing, since  $\sum_{j=1}^{k^*} \|\gamma_j\|^2 = 1$ . The coefficients in  $A_n(x)/D_n$  and  $B_n(x)/D_n$  can go either to infinity or to a constant by further selecting the subsequences of  $G_n$  and  $G'_n$ . If they go to infinity, a sequence  $d_n = O(1)$  can be found such that  $d_n A_n(x)/D_n$  converges to  $\sum_{j=1}^{k^*} \alpha_j f(x|\theta_j^*)$  and  $d_n B_n(x)/D_n$  converges to  $\sum_{j=1}^{k^*} \beta_j^T Df(x|\theta_j^*)$  for some finite  $\alpha_j$  and  $\beta_j$ . Thus, we have

$d_n$  and  $\alpha_j, \beta_j, \gamma_j$ , not all being zero, such that

$$d_n |p_{G_n}(x) - p_{G'_n}(x)| / d_\rho^2(G_n, G'_n) \rightarrow \left| \sum_{j=1}^{k^*} \alpha_j f(x|\theta_j) + \beta_j^T Df(x|\theta_j) + \gamma_j^T D^2 f(x|\theta_j) \gamma_j \right|. \quad (18)$$

for all  $x$ . This entails that the right side of the preceeding display must be 0 for all almost all  $x$ . By strong identifiability, all coefficients must be 0, which leads to contradiction.

With respect to  $\psi_1(G, G')$ , suppose that the claim is not true, which implies the existence of a subsequence  $G_n, G'_n$  such that  $\psi_1(G_n, G'_n) \rightarrow 0$ . Going through the same argument as above, we have  $\alpha_j, \beta_j, \gamma_j$ , not all of which are zero, such that Eq.(18) holds. An application of Fatou's lemma yields  $\int |\sum_{j=1}^{k^*} \alpha_j f(x|\theta_j) + \beta_j^T Df(x|\theta_j) + \gamma_j^T D^2 f(x|\theta_j) \gamma_j| d\mu = 0$ . Thus the integrand must be 0 for almost all  $x$ , leading to contradiction.  $\square$

## Proof of Theorem 2.

*Proof.* To obtain an upper bound of  $d_{\rho^2}(G, G')$  in terms of  $d_V(p_G, p_{G'})$  under the condition that  $d_V(p_G, p_{G'}) \rightarrow 0$ , our strategy is approximate  $G$  and  $G'$  by convolving these with some mollifier  $K_\delta$ . By triangular inequality,  $d_{\rho^2}(G, G')$  can be bounded in terms of  $d_{\rho^2}(G, G * K_\delta)$ ,  $d_{\rho^2}(G', G' * K_\delta)$ , and  $d_{\rho^2}(G * K_\delta, G' * K_\delta)$ . The first two terms are simple to bound, while the last term can be handled by expressing  $G * K_\delta$  as the convolution the mixture density  $p_G$  with another function. This trick was widely exploited in kernel density estimation method for deconvolution problems (e.g., Zhang (1990); Fan (1991)). We also need the following elementary lemma.

**Lemma 4.** Assume that  $p$  and  $p'$  are two probability density functions on  $\mathbb{R}^d$  with bounded  $s$ -moments.

(a) For  $t$  such that  $0 < t < s$ ,

$$\int |p(x) - p'(x)| \|x\|^t dx \leq 2 \|p - p'\|_{L_1}^{(s-t)/s} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)^{t/s}.$$

(b) Let  $V_d = \pi^{d/2} \Gamma(d/2 + 1)$  denote the volume of the  $d$ -dimensional unit sphere. Then,

$$\|p - p'\|_{L_1} \leq 2 V_d^{s/(d+2s)} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)^{\frac{d}{d+2s}} \|p - p'\|_{L_2}^{\frac{2s}{d+2s}}.$$

Take any  $s > 0$ , and let  $K : \mathbb{R}^d \rightarrow (0, \infty)$  be a symmetric density function on  $\mathbb{R}^d$  whose Fourier transform  $\tilde{K}$  is a continuous function whose support is bounded in  $[-1, 1]^d$ . Moreover,  $K$  has bounded moments up to order  $s$ . Consider molifiers  $K_\delta(x) = \frac{1}{\delta^d} K(x/\delta)$  for  $\delta > 0$ . Let  $\tilde{K}_\delta$  and  $\tilde{f}$  be the Fourier transforms for  $K_\delta$  and  $f$ , respectively. Define  $g_\delta$  to be the inverse Fourier transform of  $\tilde{K}_\delta / \tilde{f}$ :

$$g_\delta(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\langle \omega, x \rangle} \frac{\tilde{K}_\delta(\omega)}{\tilde{f}(\omega)} d\omega.$$

Note that function  $\tilde{K}_\delta(\omega)/\tilde{f}(\omega)$  has bounded support. So,  $g_\delta \in L_1(\mathbb{R})$ , and  $\tilde{g}_\delta := \tilde{K}_\delta(\omega)/\tilde{f}(\omega)$  is the Fourier transform of  $g_\delta$ . By the convolution theorem,  $f * g_\delta = K_\delta$ . As a result,

$$G * K_\delta = G * f * g_\delta = p_G * g_\delta.$$

Then the second moment under  $K_\delta$  is  $O(\delta^2)$ . It entails that  $d_{\rho^2}(G, G * K_\delta) = O(\delta^2)$ . By triangular inequality,  $d_{\rho^2}^{1/2}(G, G') \leq d_{\rho^2}^{1/2}(G * K_\delta, G' * K_\delta) + d_{\rho^2}^{1/2}(G, G * K_\delta) + d_{\rho^2}^{1/2}(G', G' * K_\delta) \leq d_{\rho^2}^{1/2}(G * K_\delta, G' * K_\delta) + O(\delta)$ , so for some constant  $C(K) > 0$  dependent only on kernel  $K$ ,

$$d_{\rho^2}(G, G') \leq 2d_{\rho^2}(G * K_\delta, G' * K_\delta) + C(K)\delta^2. \quad (19)$$

Theorem 6.15 of Villani (2008) provides an upper bound for the Wasserstein distance: for any two probability measures  $\mu$  and  $\nu$ ,  $d_{\rho^2}(\mu, \nu) \leq 2 \int \|x\|^2 d|\mu - \nu|(x)$ , where  $|\mu - \nu|$  is the total variation of measure  $|\mu - \nu|$ . Thus,

$$d_{\rho^2}(G * K_\delta, G' * K_\delta) \leq 2 \int \|x\|^2 |G * K_\delta(x) - G' * K_\delta(x)| dx. \quad (20)$$

We note that since density function  $K$  has bounded  $s$ -th moment,  $\int \|x\|^s G * K_\delta(dx) \leq 2^s [\int \|\theta\|^s dG(\theta) + \int \|x\|^s K_\delta(x) dx] = 2^s [\int \|\theta\|^s dG(\theta) + \delta^s \int \|x\|^s K(x) dx] < \infty$ , because  $G$ 's support points lie in a compact set. Applying Lemma 4 to Eq.(20), we obtain that for  $\delta < 1$ ,

$$\begin{aligned} d_{\rho^2}(G * K_\delta, G' * K_\delta) &\leq C(d, K, s) \|G * K_\delta - G' * K_\delta\|_{L_1}^{(s-2)/s} \\ &\leq C(d, K, s) \|G * K_\delta - G' * K_\delta\|_{L_2}^{2(s-2)/(d+2s)}. \end{aligned} \quad (21)$$

Here, constants  $C(d, K, s)$  are different in each line, and they are dependent only on  $d, s$  and the  $s$ -th moment of density function  $K$ .

Next, we use a known fact that for arbitrary (signed) measure  $\mu$  on  $\mathbb{R}^d$  and function  $g \in L_2(\mathbb{R}^d)$ , there holds  $\|\mu * g\|_{L_2} \leq |\mu| \|g\|_{L_2}$ , where  $|\mu|$  denotes the total variation of  $\mu$ :

$$\|G * K_\delta - G' * K_\delta\|_{L_2} = \|p_G * g_\delta - p_{G'} * g_\delta\|_{L_2} = \|(p_G - p_{G'}) * g_\delta\|_{L_2} \leq 2d_V(p_G, p_{G'}) \|g_\delta\|_{L_2}. \quad (22)$$

By Plancherel's identity,

$$\|g_\delta\|_{L_2}^2 = \frac{1}{(2\pi)^d} \int \frac{\tilde{K}_\delta(\omega)^2}{\tilde{f}(\omega)^2} d\omega = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\tilde{K}(\omega\delta)^2}{\tilde{f}(\omega)^2} d\omega \leq C \int_{[-1/\delta, 1/\delta]^d} \tilde{f}(\omega)^{-2} d\omega.$$

The last bound holds because  $\tilde{K}$  has support in  $[-1, 1]^d$ , and is bounded by a constant.

Collecting Eqs.(19)(20)(21)(22) and the preceeding display, we have:

$$d_{\rho^2}(G, G') \leq C(d, K, s) \left\{ \inf_{\delta \in (0,1)} \delta^2 + d_V(p_G, p_{G'})^{\frac{2(s-2)}{d+2s}} \left[ \int_{[-1/\delta, 1/\delta]^d} \tilde{f}(\omega)^{-2} d\omega \right]^{\frac{s-2}{d+2s}} \right\}.$$

If  $|\tilde{f}(\omega)| \prod_{j=1}^d |\omega_j|^\beta \geq d_0$  as  $\omega_j \rightarrow \infty (j = 1, \dots, d)$  for some positive constant  $d_0$ , then

$$\begin{aligned} d_{\rho^2}(G, G') &\leq C(d, K, s, \beta) \left\{ \inf_{\delta \in (0,1)} \delta^2 + d_V(p_G, p_{G'})^{\frac{2(s-2)}{d+2s}} (1/\delta)^{\frac{(2\beta+1)d(s-2)}{d+2s}} \right\} \\ &\leq C(d, K, s, \beta) d_V(p_G, p_{G'})^{\frac{4(s-2)}{2(d+2s)+(2\beta+1)d(s-2)}}. \end{aligned}$$

The exponent tends to  $4/(4 + (2\beta + 1)d)$  as  $s \rightarrow \infty$ , we obtain that  $d_{\rho^2}(G, G') \leq C(d, \beta, r) d_V(p_G, p_{G'})^r$ , for any constant  $r < 4/(4 + (2\beta + 1)d)$ , as  $d_V(p_G, p_{G'}) \rightarrow 0$ .

If  $|\tilde{f}(\omega)| \prod_{j=1}^d \exp(|\omega_j|^\beta) \geq d_0$  as  $\omega_j \rightarrow \infty (j = 1, \dots, d)$  for some positive constants  $\beta, d_0$ , then

$$d_{\rho^2}(G, G') \leq C(d, K, s, \beta) \left\{ \inf_{\delta \in (0,1)} \delta^2 + d_V(p_G, p_{G'})^{2(s-2)/(d+2s)} \exp -2d\delta^{-\beta} \frac{s-2}{d+2s} \right\}.$$

Taking  $\delta^{-\beta} = -\frac{1}{d} \log d_V(p_G, p_{G'})$ , we obtain that  $d_{\rho^2}(G, G') \leq C(d, \beta) (-\log d_V(p_G, p_{G'}))^{-2/\beta}$ .  $\square$

### Proof of Lemma 1.

*Proof.* We exploit the variational characterization of  $f$ -divergences (Nguyen et al., 2010),  $d_\phi(f_i, f'_j) = \sup_{\varphi_{ij}} \int \varphi_{ij} f'_j - \phi^*(\varphi_{ij}) f_i d\mu$ . Here, the infimum is taken over all measurable function on  $\mathcal{X}$ .  $\phi^*$  denotes the Legendre-Fenchel conjugate dual of convex function  $\phi$ . ( $\phi^*$  is again a convex function on  $\mathbb{R}$  and is defined by  $\phi^*(v) = \sup_{u \in \mathbb{R}} (uv - \phi(u))$ .) Thus,  $d_{\rho\phi}(G, G') = \inf_{\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}')} \sum_{ij} q_{ij} \sup_{\varphi_{ij}} \int \varphi_{ij} f'_j - \phi^*(\varphi_{ij}) f_i$ . On the other hand, for any  $\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}')$ ,

$$\begin{aligned} d_\phi(p_G, p_{G'}) &= \sup_{\varphi} \int \varphi p_{G'} - \phi^*(\varphi) p_G = \sup_{\varphi} \int \varphi \sum_j p'_j f'_j - \phi^*(\varphi) \sum_i p_i f_i \\ &= \sup_{\varphi} \int \varphi \sum_{ij} q_{ij} f'_j - \phi^*(\varphi) \sum_{ij} q_{ij} f_i = \sup_{\varphi} \int \sum_{ij} q_{ij} (\varphi f'_j - \phi^*(\varphi) f_i) \\ &\leq \sum_{ij} q_{ij} \sup_{\varphi_{ij}} \int (\varphi f'_j - \phi^*(\varphi) f_i), \end{aligned}$$

where the last inequality holds because the supremum is taken over a larger set of functions. Moreover, the bound holds for any  $\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}')$ , so  $d_\phi(p_G, p_{G'}) \leq d_{\rho\phi}(G, G')$ .  $\square$

### Proof of Proposition 1.

*Proof.* (a) Suppose that the claim is not true, there is a sequence of  $(G_0, G) \in \mathcal{G}_k(\Theta) \times \mathcal{G}$  such that  $d_\rho(G_0, G_2) \geq r/2 > 0$  always holds, and that converges in  $d_\rho$  metric to  $G_0^* \in \mathcal{G}_k$  and  $G^* \in \mathcal{G}$ , respectively. This is due to the compactness of both  $\mathcal{G}_k(\Theta)$  and  $\mathcal{G}$ . We must have  $d_\rho(G_0^*, G^*) \geq r/2 > 0$ , so  $G_0^* \neq G^*$ . At the same time,  $d_h(p_{G_0^*}, p_{G^*}) = 0$ , which

implies that  $p_{G_0^*} = p_{G^*}$  for almost all  $x \in \mathcal{X}$ . By finite identifiability condition,  $G_0^* = G^*$ , which is a contradiction.

(b) is an immediate consequence of Theorem 1, by noting that under the given hypothesis, there is  $c(k) > 0$  depending on  $k$ , such that  $d_h^2(p_{G_0}, p_G) \geq d_V^2(p_{G_0}, p_G)/2 \geq c(k)d_{\rho^2}^2(G_0, G) \geq c(k)d_{\rho^4}^4(G_0, G)$  for sufficiently small  $d_{\rho}(G_0, G)$ . The boundedness of  $\Theta$  implies the boundedness of  $d_{\rho}(G_0, G)$ , thereby extending the claim for the entire admissible range of  $d_{\rho}(G_0, G)$ . (c) is obtained in a similar way from Theorem 2.  $\square$

## 5.2 Strong identifiability conditions in normed spaces

Let  $(\Theta, \|\cdot\|)$  be a real normed space. A continuous function  $f : \Theta \rightarrow \mathbb{R}$  is twice Fréchet differentiable at a point  $\theta^* \in \Theta$  if it is Fréchet differentiable at  $\theta^*$ , with Fréchet derivative  $D_{\theta}f(\theta^*)$  (which is a bounded linear function from  $\Theta$  into  $\mathbb{R}$ ), and there is a continuous bilinear function  $D_{\theta}^2f(\theta^*; \cdot, \cdot)$  from  $\Theta \times \Theta$  into  $\mathbb{R}$ , called the second Fréchet differential of  $f$ , which has the property that

$$\lim_{\gamma \rightarrow 0} |f(\theta^* + \gamma) - f(\theta^*) - D_{\theta}f(\theta^*)\gamma - D_{\theta}^2f(\theta^*; \gamma, \gamma)|/\|\gamma\|^2 = 0.$$

We say that the family of density function  $\{f(\cdot|\theta), \theta \in \Theta\}$  is strongly identifiable if  $f(x|\theta)$  is twice Fréchet differentiable in  $\theta$  (with the Fréchet derivative  $D_{\theta}f(x|\theta)(\cdot)$ , and the second Fréchet differential  $D_{\theta}^2f(x|\theta; \cdot, \cdot)$ ), and for any finite  $k$  and  $k$  different  $\theta_1, \dots, \theta_k$ , the equality

$$\text{ess sup}_{x \in \mathcal{X}} \left| \sum_{i=1}^k \alpha_i f(x|\theta_i) + D_{\theta}f(x|\theta_i)\beta_i + D_{\theta}^2f(x|\theta_i; \gamma_i, \gamma_i) \right| = 0 \quad (23)$$

implies that  $\alpha_i = 0, \beta_i = \gamma_i = \mathbf{0} \in \Theta$  for  $i = 1, \dots, k$ .

Note that if for each  $x \in \mathcal{X}$ ,  $f(x|\cdot)$  is twice Fréchet continuously differentiable, i.e.,  $D_{\theta}(x|\cdot; \cdot, \cdot) : \Theta \times \Theta \times \Theta \rightarrow \mathbb{R}$  is a continuous function, then  $f$  admits the following Taylor expansion (cf. pg. 659, Polak (1997)):

$$f(x|\theta_2) - f(x|\theta_1) = D_{\theta}f(x|\theta_1)(\theta_2 - \theta_1) + \frac{1}{2}D_{\theta}^2f(x|\theta_1 + s(\theta_2 - \theta_1); (\theta_2 - \theta_1), (\theta_2 - \theta_1)),$$

for some  $s \in [0, 1]$ . Assume further that the second Fréchet differential of  $f(x|\cdot)$ ,  $D_{\theta}^2(\cdot; \cdot, \cdot)$  satisfies a uniform Lipschitz condition:

$$|D_{\theta}^2f(x|\theta_1; \gamma, \gamma) - D_{\theta}^2f(x|\theta_2; \gamma, \gamma)| \leq C\|\theta_1 - \theta_2\|^{\delta}\|\gamma\|^2$$

for all  $x, \theta_1, \theta_2 \in \Theta$ , and some fixed  $C$  and  $\delta > 0$ . It is simple to observe that Theorem 1 and its proof extend line-by-line to the normed space setting.

### 5.3 Proof of Theorem 7.

*Proof.* In the following,  $\mathbb{B}_1$  denotes the unit ball of Banach space  $\Theta = l_\infty([0, 1])$ , while  $\mathbb{H}_1$  the unit ball of the RKHS  $\mathbb{H}$ . For a given constant  $C_0 > 1$  with  $e^{-C_0 n \epsilon_n^2} < 1/2$ , define a sequence of measurable sets  $(B_n)_{n \geq 1}$ ,  $B_n = \epsilon_n \mathbb{B}_1 + C_n \mathbb{H}_1$ , where  $C_n = -2\Phi^{-1}(e^{-C_0 n \epsilon_n^2})$ . By Theorem 2.1 of van der Vaart and van Zanten (2008a), the sequence of sets  $B_n$  admits the following useful properties:

$$\log N(3\epsilon_n, B_n, \rho) \leq 6C_0 n \epsilon_n^2, \quad (24)$$

$$\Pr(\theta \notin B_n) \leq e^{-C_0 n \epsilon_n^2}, \quad (25)$$

$$\Pr(\rho(\theta, \theta_i^*) < 2\epsilon_n) \geq e^{-n \epsilon_n^2} \text{ for each } i = 1, \dots, k. \quad (26)$$

The proof proceeds by verifying that all conditions in Theorem 3 hold for some constant  $C > 0$ . Define the following sequence of subsets  $\mathcal{G}_n := \mathcal{G}_k(B_n) \subset \mathcal{G}_k(\Theta)$ . By Assumptions (B1) and (B4), there is  $c > 0$  such  $C_k(\mathcal{G}_n, r) \geq cr^4$  for sufficiently small  $r \geq 0$ .

Let  $C_4 = \mathbb{E}\|\theta\|^2 < \infty$ , then for any  $h \in \mathbb{H}$ ,  $\|h\| \leq C_4 \|h\|_{\mathbb{H}}$  (cf. van der Vaart and van Zanten (2008a) pg. 203). Thus,  $\text{Diam}(B_n) \leq 2(\epsilon_n + C_4 C_n) \leq 2(\epsilon_n + C_4 \epsilon_n \sqrt{10C_0 n})$  (cf. van der Vaart and van Zanten (2008b) pg. 1454, for the second equality). By Lemma 2(a),  $\log N(2\epsilon_n, \mathcal{G}_n, d_\rho) \leq k(\log N(\epsilon_n, B_n, \rho) + \log(e + e \text{Diam}(B_n)/\epsilon_n))$ . Combined with (24), we have  $\log D(4\epsilon_n, \mathcal{G}_n, d_\rho) \leq \log N(2\epsilon_n, \mathcal{G}_n, d_\rho) \leq O(kn\epsilon_n^2)$ , due to (B4) and the assumption that  $\log n = o(n\epsilon_n^2)$ . If we replace  $\epsilon_n$  by a sufficiently large multiple of  $\epsilon_n$ , we shall obtain the bound (7) precisely.

Turning to (8), by the union bound,  $\Pr(G \notin \mathcal{G}_n) \leq \sum_{i=1}^k \Pr(\theta_i \notin B_n) \leq ke^{-C_0 n \epsilon_n^2} \leq \exp(-n\epsilon_n^2(C+4))$ , by choosing constant  $C_0$  sufficiently large (after  $C$  is fixed).

Next, we consider condition (9). Suppose that  $G = \sum_{i=1}^k p_i \delta_{\theta_i^*}$  where  $\rho(\theta_i, \theta_i^*) \leq \epsilon_n$  for all  $i = 1, \dots, k$ , and  $|p_i - p_i^*| \leq \epsilon_n^2/(k \text{Diam}(B_n)^2)$ . Combining Lemma 1 with Assumption (B2) on the likelihood functions, we obtain that  $d_K(p_{G_0}, p_G) \leq d_{\rho K}(G_0, G) \leq C_1 \sum_{1 \leq i, j \leq k} q_{ij} \rho^2(\theta_i^*, \theta_j)$ , for any  $\mathbf{q} \in \mathcal{Q}$ . It is simple to check that  $\inf_{\mathbf{q}} \sum_{1 \leq i, j \leq k} q_{ij} \rho^2(\theta_i^*, \theta_j) \leq \sum_{i=1}^k (p_i^* \wedge p_i) \rho^2(\theta_i^*, \theta_i) + \sum_{i=1}^k |p_i - p_i^*| \text{Diam}(B_n)^2 \leq 2\epsilon_n^2$ . Hence,

$$\begin{aligned} \Pi(d_K(p_{G_0}, p_G) \leq 2\epsilon_n^2) &\geq \Pi(\rho(\theta_i, \theta_i^*) \leq \epsilon_n; |p_i - p_i^*| \leq \epsilon_n^2/(k \text{Diam}(B_n)^2), i = 1, \dots, k) \\ &\geq \exp(-kn\epsilon_n^2/4) c_3 (\epsilon_n^2/(k \text{Diam}(B_n)^2))^{k\alpha} \\ &\geq c_3 \exp(-kn\epsilon_n^2/4) \left( 4k(1 + C_4 \sqrt{10C_0 n})^2 \right)^{-k\alpha} \\ &\geq \exp(-n\epsilon_n^2 C). \end{aligned}$$

(The second inequality is due to Assumption (B3) and (26), the fourth inequality is due to Assumption (B4)). In view of Assumption (B2), we obtain that condition (9) holds by choosing a sufficiently large constant  $C$ .

Finally, we shall choose  $M_n$  such that  $M_n \epsilon_n \rightarrow 0$ , and  $C_k(\mathcal{G}_n, M_n \epsilon_n) \geq c(M_n \epsilon_n)^4 \geq \epsilon_n^2(C+4)$ . This is possible by taking  $M_n$  to be a large multiple of  $\epsilon_n^{-1/2}$ . As a result,  $M_n \epsilon_n \asymp \epsilon_n^{1/2}$ . We conclude by invoking Thm 3.  $\square$

## 5.4 Other auxiliary results

### Proof of Lemma 2.

*Proof.* (a) Suppose that  $(\eta_1, \dots, \eta_T)$  forms an  $\epsilon$ -covering for  $\Theta$  under metric  $\rho$ , where  $T = N(\epsilon, S, \rho)$  denote the (minimum) covering number. Take any discrete measure  $G(\mathbf{p}, \boldsymbol{\theta}) = \sum_{i=1}^k p_i \delta_{\theta_i}$ . For each  $\theta_i$  there is an approximating  $\theta'_i$  among the  $\eta_j$ 's such that  $\rho(\theta_i, \theta'_i) < \epsilon$ . Let  $\mathbf{p}' = (p'_1, \dots, p'_k)$  be a  $k$ -dim vector in the probability simplex that deviates from  $\mathbf{p}$  by less than  $\delta$  in  $l_1$  distance:  $\|\mathbf{p}' - \mathbf{p}\|_1 \leq \delta$ . Define  $G' = \sum_{i=1}^k p'_i \delta_{\theta'_i}$ . Then  $d_\rho(G, G') \leq \sum_{i=1}^k (p_i \wedge p'_i) \rho(\theta_i, \theta'_i) + \|\mathbf{p} - \mathbf{p}'\|_1 \text{Diam}(\Theta) \leq \epsilon + \delta \text{Diam}(\Theta)$ . (This is easily seen by moving  $p_i \wedge p'_i$  mass of  $\theta_i$  to the “nearby”  $\theta'_i$ , while the remaining mass are moved in arbitrary way). It follows that a  $(\epsilon + \delta \text{Diam}(\Theta))$ -covering for  $\mathcal{G}_k(\Theta)$  can be constructed by combining each element of a  $\delta$ -covering in  $l_1$  metric of the  $k - 1$ -probability simplex and  $k$   $\epsilon$ -covering's of  $\Theta$ .

The covering number of  $k - 1$ -probability simplex is less than the number of cubes of length  $\delta/k$  covering  $[0, 1]^k$  times the volume of  $\{(p'_1, \dots, p'_k) : p'_j \geq 0, \sum_j p'_j \leq 1 + \delta\}$ , i.e.,  $(k/\delta)^k (1 + \delta)^k / k! \sim (1 + 1/\delta)^k e^k / \sqrt{2\pi k}$ . It follows that  $N(\epsilon + \delta \text{Diam}(\Theta), \mathcal{G}_k(\Theta), d_\rho) \leq T^k (1 + 1/\delta)^k e^k / \sqrt{2\pi k}$ . Take  $\delta = \epsilon / \text{Diam}(\Theta)$  to achieve the claim.

(b) Suppose that  $(\eta_1, \dots, \eta_T)$  forms an  $\epsilon$ -covering for  $\Theta$  under metric  $\rho$ , and  $T = N(\epsilon, S, \rho)$ . Take any discrete measure  $G(\mathbf{p}, \boldsymbol{\theta}) = \sum_{i=1}^k p_i \delta_{\theta_i} \in \bar{\mathcal{G}}_\Theta$ , where  $k$  may be infinity. The collection of atoms  $\theta_1, \dots, \theta_k$  can be subdivide into disjoint subsets  $S_1, \dots, S_T$ , some of which may be empty, so that for each  $t = 1, \dots, T$ ,  $\rho(\theta_i, \eta_t) \leq \epsilon$  for any  $\theta_i \in S_t$ . Thus, if we define  $p'_t = \sum_{i=1}^k p_i \mathbb{I}(\theta_i \in S_t)$ , and discrete measure  $G'(\mathbf{p}', \boldsymbol{\eta}) = \sum_{t=1}^T p'_t \delta_{\eta_t}$ , then we are guaranteed that  $d_\rho(G, G') \leq \sum_{i=1}^k \sum_{t=1}^T p_i \mathbb{I}(\theta_i \in S_t) \rho(\theta_i, \eta_t) \leq \epsilon$ .

Let  $\mathbf{p}'' = (p''_1, \dots, p''_T)$  be a  $T$ -dim vector in the probability simplex that deviates from  $\mathbf{p}'$  by less than  $\delta$  in  $l_1$  distance:  $\|\mathbf{p}'' - \mathbf{p}'\|_1 \leq \delta$ . Take  $G'' = \sum_{t=1}^T p''_t \delta_{\eta_t}$ . It is simple to observe that  $d_\rho(G', G'') \leq \text{Diam}(\Theta) \delta$ . By triangle inequality,  $d_\rho(G, G'') \leq d_\rho(G, G') + d_\rho(G', G'') \leq \epsilon + \delta \text{Diam}(\Theta)$ .

The foregoing arguments establish that  $(\epsilon + \delta \text{Diam}(\Theta))$ -covering in the Wasserstein metric for the subset  $\mathcal{G}_S \subseteq \mathcal{G}(\Theta)$  can be constructed by combining each element of the  $\delta$ -covering in  $l_1$  of the  $T - 1$  simplex and a single covering of  $\Theta$ . From the proof of part (a),  $N(\epsilon + \delta \text{Diam}(\Theta), \mathcal{G}_S, d_\rho) \leq (1 + 1/\delta)^T e^T / \sqrt{2\pi T}$ . Take  $\delta = \epsilon / \text{Diam}(\Theta)$  to conclude.

(c) Consider a  $G = \sum_{i=1}^k p_i \delta_{\theta_i}$  such that  $d_\rho(G_0, G) \leq 2\epsilon$ . By definition, there is  $q \in \mathcal{Q}(\mathbf{p}, \mathbf{p}^*)$  so that  $\sum_{ij} q_{ij} \rho(\theta_i^*, \theta_j) \leq 2\epsilon$ . Since  $\sum_j q_{ij} = p_i^*$ , this implies that  $2\epsilon \geq \sum_{i=1}^k p_i^* \min_j \rho(\theta_i^*, \theta_j)$ . Thus, for each  $i = 1, \dots, k$  there is a  $j$  such that  $\rho(\theta_i^*, \theta_j) \leq 2\epsilon / p_i^* \leq 2M\epsilon$ . Without loss of generality, assume that  $\rho(\theta_i^*, \theta_i) \leq 2M\epsilon$  for all  $i = 1, \dots, k$ . For sufficiently small  $\epsilon$ , for any  $i$ , it is simple to observe that  $d_\rho(G_0, G) \geq |p_i^* - p_i| \min_{j \neq i} \rho(\theta_i^*, \theta_j) \geq |p_i^* - p_i| \min_j \rho(\theta_i^*, \theta_j^*) / 2$ . Thus,  $|p_i^* - p_i| \leq 4\epsilon / m$ .

Thus, an  $\epsilon/4 + \delta \text{Diam}(\Theta)$  covering in  $d_\rho$  for  $\{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}$  can be constructed by combining the  $\epsilon/4$ -covering for each of the  $k$  sets  $\{\theta \in \Theta : \rho(\theta, \theta_i^*) \leq 2M\epsilon\}$  and the  $\delta/k$ -covering for each of the  $k$  sets  $[p_i^* - 4\epsilon/m, p_i^* + 4\epsilon/m]$ . This entails that:  $N(\epsilon/4 + \delta \text{Diam}(\Theta), \{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq [\sup_{\Theta'} N(\epsilon/4, \Theta', \rho)]^k (8\epsilon k/m\delta)^k$ .

Take  $\delta = \epsilon/(4\text{Diam}(\Theta))$  to conclude the proof.  $\square$

#### Proof of Lemma 4.

*Proof.* (a) For arbitrary constant  $R > 0$ , we have  $\int |p(x) - p'(x)| \|x\|^t dx \leq \int_{\|x\| \leq R} |p - p'| \|x\|^t + \int_{\|x\| \geq R} (p + p') \|x\|^t \leq R^t \|p - p'\|_{L_1} + R^{-(s-t)} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)$ . Choosing  $R = [(\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s) / \|p - p'\|_{L_1}]^{1/s}$  to conclude.

(b) For any  $R > 0$ , we have  $\int_{\|x\| \leq R} |p(x) - p'(x)| dx \leq V_d^{1/2} R^{d/2} [\int_{\|x\| \leq R} (p(x) - p'(x))^2 dx]^{1/2} \leq V_d^{1/2} R^{d/2} \|p - p'\|_{L_2}$ . We also have  $\int_{\|x\| \geq R} |p(x) - p'(x)| dx \leq \int_{\|x\| \geq R} p(x) + p'(x) dx \leq R^{-s} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)$ . Thus,  $\|p - p'\|_{L_1} \leq \inf_{R>0} V_d^{1/2} R^{d/2} \|p - p'\|_{L_2} + R^{-s} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)$ , which gives the desired bound.  $\square$

## 6 Appendix

We outline in this section the proofs of theorems 3 and 4 for completeness. Our proof follows the same steps as in Ghosal et al. (2000), with suitable modification for the inclusion of the Hellinger information function, which plays important roles in the specification of conditions for convergence and determining convergence rates. The proof consists of results on the existence of test, which is then turned into probability bounds on the posterior contraction.

A test  $\phi_n$  is a measurable indicator function of the iid sample  $X_1, \dots, X_n$ . For each pair of discrete measures  $G_0, G_1$  we consider tests for discriminating  $G_0 \in \mathcal{G}(\Theta)$  against a closed ball  $B(G_1, d_\rho(G_0, G_1)/2) = \{G \in \bar{\mathcal{G}}(\Theta) : d_\rho(G_1, G) \leq d_\rho(G_1, G_0)/2\}$ . In the following  $P_G$  denotes the expectation under the mixture distribution given by density  $p_G$ .

**Lemma 5.** *For some fixed  $k < \infty$ , suppose that  $\mathcal{G}_k(\Theta) \subseteq \mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$ . Then, for every pair of discrete measures  $(G_0, G_1) \in \mathcal{G}_k(\Theta) \times \mathcal{G}$  there exist tests  $\{\phi_n\}$  that have the following properties:*

$$P_{G_0} \phi_n \leq \exp[-nC_k(\mathcal{G}, d_\rho(G_0, G_1))] \quad (27)$$

$$\sup_{G \in \bar{\mathcal{G}}(\Theta) : d_\rho(G, G_1) < d_\rho(G_0, G_1)/2} P_G(1 - \phi_n) \leq \exp[-nC_k(\mathcal{G}, d_\rho(G_0, G_1))]. \quad (28)$$

Next, existence of test can be shown for discriminating  $G_0$  against the complement of a closed ball:

**Lemma 6.** *Let  $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$ .  $G_0 \in \mathcal{G}_k(\Theta) \subseteq \mathcal{G}$  for some  $k < \infty$ . Suppose that for some non-increasing function  $D(\epsilon)$ , some  $\epsilon_n \geq 0$  and every  $\epsilon > \epsilon_n$ ,*

$$D(\epsilon/2, \{G \in \mathcal{G} : \epsilon \leq d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq D(\epsilon). \quad (29)$$



Then, for every  $\epsilon > \epsilon_n$  there exist tests  $\varphi_n$  (depending on  $\epsilon > 0$ ) such that for any  $t \in \mathbb{N}$

$$P_{G_0} \varphi_n \leq D(\epsilon) \sum_{t=1}^{\lceil \text{Diam}(\Theta)/\epsilon \rceil} \exp[-nC_k(\mathcal{G}, t\epsilon)] \quad (30)$$

$$\sup_{G \in \mathcal{G}: d_\rho(G_0, G) > t\epsilon} P_G(1 - \varphi_n) \leq \exp[-nC_k(\mathcal{G}, t\epsilon)]. \quad (31)$$

**Proof of Theorem 3 and 4** By a result of Ghosal et al (Ghosal et al., 2000) (Lemma 8.1, pg. 524), for every  $\epsilon > 0$  and probability measure  $\Pi$  on the set  $B(\epsilon)$  defined by Eq. (6), we have, for every  $C > 0$ ,

$$P_{G_0} \left( \int \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} d\Pi(G) \leq \exp(-(1+C)n\epsilon^2) \right) \leq \frac{1}{C^2 n \epsilon^2}.$$

This entails that, for a fixed  $C \geq 1$ , there is an event  $A_n$  with  $P_{G_0}$ -probability at least  $1 - (Cn\epsilon_n^2)^{-1}$ , for which there holds:

$$\int \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i) d\Pi(G) \geq \exp(-2n\epsilon_n^2) \Pi(B(\epsilon_n)). \quad (32)$$

Let  $\mathcal{O}_n = \{G \in \bar{\mathcal{G}}(\Theta) : d_\rho(G_0, G) \geq M_n \epsilon_n\}$ ,  $S_{n,j} = \{G \in \mathcal{G}_n : d_\rho(G_0, G) \in [j\epsilon_n, (j+1)\epsilon_n)\}$  for each  $j \geq 1$ . The conditions specified by Lemma 6 are satisfied by setting  $D(\epsilon) = \exp(n\epsilon_n^2)$  (constant in  $\epsilon$ ). Thus there exist tests  $\varphi_n$  for which Eq. (30) and (31) hold. Then,

$$\begin{aligned} & P_{G_0} \Pi(G \in \mathcal{O}_n | X_1, \dots, X_n) \\ &= P_{G_0} [\varphi_n \Pi(G \in \mathcal{O}_n | X_1, \dots, X_n)] + P_{G_0} [(1 - \varphi_n) \Pi(G \in \mathcal{O}_n | X_1, \dots, X_n)] \\ &\leq P_{G_0} [\varphi_n \Pi(G \in \mathcal{O}_n | X_1, \dots, X_n)] + P_{G_0} \mathbb{I}(A_n^c) + P_{G_0} [(1 - \varphi_n) \Pi(G \in \mathcal{O}_n | X_1, \dots, X_n) \mathbb{I}(A_n)]. \end{aligned}$$

Exploiting Lemma 6, all terms in the preceeding display can be shown to vanish as  $n \rightarrow \infty$ . The proof for Theorem 3 proceeds in a similar way to Theorem 2.1 of Ghosal et al. (2000), while the proof for Theorem 4 is similar to their Theorem 2.4.

## References

- Barron, A., Schervish, M., and Wasserman, L. (1999), “The consistency of posterior distributions in nonparametric problems,” *Ann. Statist.*, 27, 536–561.
- Bickel, P. and Freedman, D. (1981), “Some asymptotic theory for the bootstrap,” *Annals of Statistics*, 9, 1196–1217.
- Chen, J. (1995), “Optimal rate of convergence for finite mixture models,” *Annals of Statistics*, 23, 221–233.

- del Barrio, E., Cuesta-Albertos, J., Matrán, C., and Rodríguez-Rodríguez, J. (1999), “Tests of goodness of fit based on the  $L_2$ -Wasserstein distance,” *Annals of Statistics*, 27, 1230–1239.
- Dudley, R. M. (1976), *Probabilities and metrics: Convergence of laws on metric spaces, with a view to statistical testing*, Aarhus Universitet.
- Fan, J. (1991), “On the optimal rates of convergence for nonparametric deconvolution problems,” *Annals of Statistics*, 19, 1257–1272.
- Ferguson, T. (1973), “A Bayesian analysis of some nonparametric problems,” *Ann. Statist.*, 1, 209–230.
- Gelfand, A., Kottas, A., and MacEachern, S. (2005), “Bayesian nonparametric spatial modeling with Dirichlet process mixing,” *J. Amer. Statist. Assoc.*, 100, 1021–1035.
- Genovese, C. and Wasserman, L. (2000), “Rates of convergence for the Gaussian mixture sieve,” *Annals of Statistics*, 28, 1105–1127.
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), “Posterior consistency of Dirichlet mixtures in density estimation,” *Ann. Statist.*, 27, 143–158.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. (2000), “Convergence rates of posterior distributions,” *Ann. Statist.*, 28, 500–531.
- Ghosal, S. and van der Vaart, A. (2007a), “Convergence rates of posterior distributions for noniid observations,” *Ann. Statist.*, 35, 192–223.
- (2007b), “Posterior convergence rates of Dirichlet mixtures at smooth densities,” *Ann. Statist.*, 35, 697–723.
- Hjort, N., Holmes, C., Mueller, P., and Walker, S. (2010), *Bayesian Nonparametrics: Principles and Practice*, Cambridge University Press.
- Ishwaran, H., James, L., and Sun, J. (2001), “Bayesian model selection in finite mixtures by marginal density decompositions,” *Journal of American Statistical Association*, 96, 1316–1332.
- Ishwaran, H. and Zarepour, M. (2002), “Dirichlet prior sieves in finite normal mixtures,” *Statistica Sinica*, 12, 941–963.
- Kallianpur, G. (1971), “Abstract Wiener processes and their reproducing kernel Hilbert spaces,” *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 17, 113–123.
- Lindsay, B. (1995), *Mixture models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA.

- Mallows, C. (1972), “A note on asymptotic joint normality,” *Annals of Mathematical Statistics*, 43, 508–515.
- McLachlan, G. and Basford, K. (1988), *Mixture models: Inference and Applications to Clustering*, New York: Marcel-Dekker.
- Nguyen, X. (2010), “Inference of global clusters from locally distributed data,” *Bayesian Analysis*, 5, 817–846.
- Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010), “Estimating divergence functionals and the likelihood ratio by convex risk minimization,” *IEEE Transactions on Information Theory*, 56, 5847–5861.
- Petrone, S., Guidani, M., and Gelfand, A. (2009), “Hybrid Dirichlet processes for functional data,” *Journal of the Royal Statistical Society B*, 71(4), 755–782.
- Polak, E. (1997), *Optimization: Algorithms and Consistent Approximations*, Springer.
- Rodriguez, A., Dunson, D., and Gelfand, A. (2008), “The nested Dirichlet process,” *J. Amer. Statist. Assoc.*, 103, 1131–1154.
- Shen, X. and Wasserman, L. (2001), “Rates of convergence of posterior distributions,” *Ann. Statist.*, 29, 687–714.
- Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), “Hierarchical Dirichlet processes,” *J. Amer. Statist. Assoc.*, 101, 1566–1581.
- Teicher, H. (1960), “On the mixture of distributions,” *Ann. Math. Statist.*, 31, 55–73.
- (1961), “Identifiability of mixtures,” *Ann. Math. Statist.*, 32, 244–248.
- van der Vaart, A. and van Zanten, J. (2008a), “Rates of contraction of posterior distributions based on Gaussian process priors,” *Annals of Statistics*, 36, 1435–1463.
- (2008b), “Reproducing kernel Hilbert spaces of Gaussian priors,” *IMS Collections: Pushing the Limits of Contemporary Statistics: Contributions in Honors of Jayanta K. Ghosh*, 3, 200–222.
- Villani, C. (2003), *Topics in Optimal Transportation*, American Mathematical Society.
- (2008), *Optimal transport: Old and New*, Springer.
- Walker, S. (2004), “New approaches to Bayesian consistency,” *Ann. Statist.*, 32, 2028–2043.
- Walker, S., Lijoi, A., and Prunster, I. (2007), “On rates of convergence for posterior distributions in infinite-dimensional models,” *Ann. Statist.*, 35, 738–746.
- Zhang, C. (1990), “Fourier methods for estimating mixing densities and distributions,” *Annals of Statistics*, 18, 806–831.